Aspects of Noncommutative Harmonic Analysis

Jean Gallier and Jocelyn Quaintance
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
e-mail: jean@cis.upenn.edu

© Jean Gallier

March 17, 2018
Contents

1 Function Spaces Often Encountered ........................................... 7
   1.1 Spaces of Bounded Functions ............................................ 7
   1.2 Simple and Uniform Convergence ....................................... 9
   1.3 Regulated Functions ......................................................... 15

2 The Riemann Integral ................................................................. 23
   2.1 Riemann Integral of a Continuous Function ......................... 24
   2.2 The Riemann Integral of Regulated Functions ....................... 29

3 Measure Theory; Basic Notions ................................................ 33
   3.1 σ-Algebras, Measures ...................................................... 34
   3.2 Null Subsets and Properties Holding Almost Everywhere .......... 43
   3.3 Construction of a Measure from an Outer Measure ................. 45
   3.4 The Lebesgue Measure on \( \mathbb{R} \) .................................. 48

4 Integration .................................................................................. 53
   4.1 Measurable Maps .................................................................. 55
   4.2 Step Maps on a Measurable Space .................................... 59
   4.3 \( \mu \)-Measurable Maps and \( \mu \)-Step Maps ......................... 61
   4.4 The Integral of \( \mu \)-Step Maps ........................................ 68
   4.5 Integrable Functions; the Spaces \( \mathcal{L}_\mu(X, \mathcal{A}, F) \) and \( L_\mu(X, \mathcal{A}, F) \) .......... 73
   4.6 Fundamental Convergence Theorems .................................. 83
   4.7 The Spaces \( \mathcal{L}_\mu^p(X, \mathcal{A}, F) \) and \( L_\mu^p(X, \mathcal{A}, F) \); \( p = 1, 2, \infty \) ....... 90
   4.8 Products of Measure Spaces and Fubini’s Theorem ............... 96

5 Radon Measures on Locally Compact Spaces ................................ 103
   5.1 Positive Radon Functionals Induced by Borel Measures .......... 104
   5.2 The Radon–Riesz Theorem and Positive Radon Functionals .... 109
   5.3 Regular Borel Measures ...................................................... 112
   5.4 Complex and Real Measures .............................................. 115
   5.5 Real Measures and the Hahn–Jordan Decomposition .......... 118
   5.6 Total Variation of a Radon Functional ................................ 120
### Contents

#### 5.7 The Radon–Riesz Theorem and Bounded Radon Functionals

#### 6 The Haar Measure
- 6.1 Topological Groups
- 6.2 Existence of the Haar Measure
- 6.3 Uniqueness of the Haar Measure
- 6.4 The Modular Function
- 6.5 Examples of Haar Measures
- 6.6 Haar Measure on a Quotient Group
- 6.7 Convolution

#### 7 The Fourier Transform on \( \mathbb{R}^n \)

#### 8 Normed Algebras and Spectral Theory
- 8.1 Normed Algebras
- 8.2 Spectrum in a Normed Algebra
- 8.3 Characters and Spectrum in a Normed Algebra
- 8.4 Banach Algebras with Involution
- 8.5 Representations of Banach Algebras with Involution
- 8.6 Positive Linear Forms and Positive Hilbert Forms
- 8.7 Traces, Bitraces, Hilbert Algebras
- 8.8 Complete Hilbert Algebras
- 8.9 The Plancherel–Godement Theorem

#### 9 Representations of Locally Compact Groups
- 9.1 Group Representations
- 9.2 Representations of \( G \) and Representations of \( L^1(G) \)
- 9.3 Functions of Positive Type

#### 10 Locally Compact Abelian Groups
- 10.1 Characters
- 10.2 The Dual Group
- 10.3 The Fourier Transform
- 10.4 Pontrjagin Duality
- 10.5 Spectral Synthesis

#### 11 Representations of Compact Groups
- 11.1 The Peter–Weyl Theorem
- 11.2 The Fourier Transform

#### 12 Gelfand Pairs
- 12.1 Characters
- 12.2 The Fourier Transform
CONTENTS

13 Topology 163
  13.1 Metric Spaces and Normed Vector Spaces 163
  13.2 Topological Spaces 170
  13.3 Continuous Functions, Limits 180
  13.4 Connected Sets 187
  13.5 Compact Sets and Locally Compact Spaces 197
  13.6 Second-Countable and Separable Spaces 209
  13.7 Sequential Compactness 213
  13.8 Complete Metric Spaces and Compactness 219
  13.9 Completion of a Metric Space 222
  13.10 The Contraction Mapping Theorem 229
  13.11 Continuous Linear and Multilinear Maps 233
  13.12 Completion of a Normed Vector Space 240
  13.13 Further Readings 242

14 Vector Norms and Matrix Norms 243
  14.1 Normed Vector Spaces 243
  14.2 Matrix Norms 249

15 Basics of Groups and Group Actions 263
  15.1 Groups, Subgroups, Cosets 263
  15.2 Group Actions: Part I, Definition and Examples 276
  15.3 Group Actions: Part II, Stabilizers and Homogeneous Spaces 289
  15.4 The Grassmann and Stiefel Manifolds 297

Bibliography 303
Chapter 1

Function Spaces Often Encountered

Various spaces of functions \( f : E \to F \) from a topological space \( E \) to a metric space or a normed vector space \( F \) come up all the time. The most frequently encountered are the spaces \( (F^E)_b \) of bounded functions, the spaces \( \mathcal{K}(E; F) \) of continuous functions with compact support, the spaces \( \mathcal{C}_0(E; F) \) of continuous functions which tend to zero at infinity, and the spaces \( \mathcal{C}_b(E; F) \) of bounded continuous functions. When \( F \) is a normed vector space, all these spaces are normed vector spaces with the sup norm. An important issue about function spaces is the convergence of sequences of functions. We review the main two notions, pointwise convergence (also known as simple convergence) and uniform convergence. A sequence of continuous functions may converge pointwise to a function which is not continuous. Uniform convergence has a better behavior. If \( F \) is a complete normed vector space, then both spaces \( \mathcal{C}_b(E; F) \) and \( (F^E)_b \) are also complete. An interesting family of functions in \( (F^{[a,b]})_b \) is the space \( \text{Reg}([a,b]; F) \) of regulated functions. These functions only have at most countably many simple kinds of discontinuities called discontinuities of the first kind. If \( F \) is a complete normed vector space, then the space \( \text{Step}([a,b]; F) \) is complete. It contains a subspace \( \text{Step}([a,b]; F) \) consisting of very simple functions called step functions, which take finitely many different values on consecutive open intervals. The space \( \text{Step}([a,b]; F) \) is dense in \( \text{Reg}([a,b]; F) \). If \( E \) is a locally compact space, then the space \( \mathcal{C}_0(E; \mathbb{C}) \) is the closure of \( \mathcal{K}_C(E) \) in \( \mathcal{C}_b(E; \mathbb{C}) \).

1.1 Spaces of Bounded Functions

In this section, we are dealing with functions \( f : E \to F \), where \( F \) is either a metric space or a normed vector space. Recall that the set of all functions from \( E \) to \( F \) is denoted by \( F^E \).

First assume that \( F \) is a metric space with metric \( d \). We would like to make \( F^E \) into a metric space. It is natural to define a metric on \( F^E \) by setting

\[
d_\infty(f, g) = \sup_{x \in E} d(f(x), g(x))
\]

for any two functions \( f, g : E \to F \), but if \( d(f(x), g(x)) \) is unbounded as \( x \) ranges over \( E \), the
expression \( \sup_{x \in E} d(f(x), g(x)) \) is undefined. Therefore, we consider the space of bounded functions defined as follows.

**Definition 1.1.** If \((F, d)\) is a metric space, a function \(f : E \to F\) is bounded if its image \(f(E)\) is bounded in \(F\), which means that \(f(E) \subseteq B(a, \alpha)\), for some closed ball \(B(a, \alpha)\) of center \(a\) and radius \(\alpha > 0\). See Figure 1.1. The space of bounded functions \(f : E \to F\) is denoted by \((F^E)_b\).

![Figure 1.1](image)

Figure 1.1: Let \(E = F = \mathbb{R}\) with the Euclidean metric. In Figure (i), \(f\) is unbounded since \(f(E) = \mathbb{R}\). In Figure (ii), \(f \in (F^E)_b\) since \(f(E) = (0, 1]\) and \((0, 1] \subset B(0, 1) = [-1, 1]\).

If \(f : E \to F\) and \(g : E \to F\) are bounded functions then it is easy to see that if \(f(E) \subseteq B(a, \alpha)\) and if \(g(E) \subseteq B(b, \beta)\), then
\[
d(f(x), g(x)) \leq \alpha + \beta + d(a, b) \quad \text{for all } x \in E.
\]
Therefore, \(\sup_{x \in E} d(f(x), g(x))\) is well defined. It is easy to check that if we define
\[
d_\infty(f, g) = \sup_{x \in E} d(f(x), g(x))
\]
for any two bounded functions \(f, g\), then \(d\) is indeed a metric on \((F^E)_b\).

**Definition 1.2.** If \((F, d)\) is a metric space, then for any two bounded functions \(f, g \in (F^E)_b\), the quantity
\[
d_\infty(f, g) = \sup_{x \in E} d(f(x), g(x))
\]
is a metric on \((F^E)_b\). See Figure 1.2.

If \((F, \| \cdot \|)\) is normed metric space, then \(F^E\) is a vector space, and it is easy to check that \((F^E)_b\) is also a vector space. For any bounded function \(f : E \to F\) (which means that \(f(E) \subseteq B(0, \alpha)\), for some closed ball \(B(0, \alpha)\)), then
\[
\|f\|_\infty = \sup_{x \in E} \|f(x)\|
\]
is a norm on the vector space \((F^E)_b\).
1.2. SIMPLE AND UNIFORM CONVERGENCE

Figure 1.2: Let $E = F = \mathbb{R}$ with the Euclidean metric. Both $f, g \in (F^E)_b$ since $f(E) = (0, 1)$, while $g(E) = [-1, 0)$. The concatenation of the vertical dashed red lines is $d_\infty(f, g) = \sup_{x \in E} d(f(x), g(x)) = 1 - (-1) = 2$.

**Definition 1.3.** If $(F, \|\|)$ is a normed vector space, then for any bounded function $f \in (F^E)_b$, the quantity

$$\|f\|_\infty = \sup_{x \in E} \|f(x)\|$$

is a norm on $(F^E)_b$, often called the sup norm.

The following important theorem can be shown; see Schwartz [38] (Chapter XV, Section 1, Theorem 1).

**Theorem 1.1.**  
(1) If $(F, d)$ is a complete metric space, then $((F^E)_b, d_\infty)$ is also a complete metric space.

(2) If $(F, \|\|)$ is a complete normed vector space, then $((F^E)_b, \|\|_\infty)$ is also a complete normed vector space.

1.2 Simple and Uniform Convergence

When dealing with spaces of functions, a crucial issue is to identify notions of limit that preserve certain desirable properties, such as continuity. There are primarily two such notions, simple and uniform convergence that we now review.

**Definition 1.4.** Let $(F, d)$ be a metric space. A sequence $(f_n)_{n \geq 1}$ of functions $f_n : E \to F$ converges simply (or converges pointwise) to a function $f : E \to F$ if for every $x \in E$, for every $\epsilon > 0$, there is some $N > 0$ such that

$$d(f_n(x), f(x)) < \epsilon \quad \text{for all} \ n \geq N.$$ 

See Figure 1.3.
CHAPTER 1. FUNCTION SPACES OFTEN ENCOUNTERED

Figure 1.3: A schematic illustration of $f_n(x)$ converging pointwise $f(x)$, where $E = F = \mathbb{R}$. As $n$ increases, the graph of $f_n(x)$ near $x$ must be in the band determined by the graphs of $f(x) - \epsilon$ and $f(x) + \epsilon$.

Definition 1.4 says that for every $x \in E$, the sequence $(f_n(x))_{n \geq 1}$ converges to $f(x)$. Observe that the above $\epsilon$ depends on $x$.

Remark: The product topology on $F^E$ is defined as follows: a subset of functions in $F^E$ is open if it is the union of subsets $U_A$ of functions $f : E \to F$ for which there is some finite subset $A$ of $E$ such that $f(x) \in U_x$ for all $x \in A$, where $U_x$ is an open subset of $F$, and $f(x) \in F$ is arbitrary for all $x \in E - A$. Then it is easy to see that a sequence $(f_n)_{n \geq 1}$ of elements of $F^E$ converges simply to $f \in F^E$ iff the sequence $(f_n)_{n \geq 1}$ converge to $f$ in the product topology. This is sometimes refered to as weak convergence.

Definition 1.5. Let $(F,d)$ be a metric space. A sequence $(f_n)_{n \geq 1}$ of functions $f_n : E \to F$ converges uniformly to a function $f : E \to F$ if for every $\epsilon > 0$, there is some $N > 0$ such that

$$d(f_n(x), f(x)) < \epsilon \quad \text{for all } n \geq N \text{ and for all } x \in E.$$ 

See Figure 1.4.

The difference between simple and uniform convergence is that in uniform convergence, $\epsilon$ is independent of $x$. For example the functions $f_n : [0, 2\pi] \to \mathbb{R}$ defined by $f_n(x) = n \sin \left( \frac{x}{n} \right)$ converges uniformly to $f(x) = x$, as evidenced by Figure 1.5. Consequently, uniform convergence implies simple convergence, but the converse is false, as the following examples illustrate.

Example 1.1. Let $g : \mathbb{R} \to \mathbb{R}$ be the function given by

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

and for every $n \geq 1$, let $f_n : \mathbb{R} \to \mathbb{R}$ be the function given by

$$f_n(x) = \frac{1}{1 + (x - n)^2}.$$
1.2. SIMPLE AND UNIFORM CONVERGENCE

Figure 1.4: A schematic illustration of $f_n$ converging uniformly to $f$, where $E = F = \mathbb{R}$. As $n$ increases, the graph of $f_n$ must lie entirely in the band determined by the graphs of $f - \epsilon$ and $f + \epsilon$.

The function $f_n$ is obtained by translating $g$ to the right using the translation $x \mapsto x + n$; see Figure 1.6.

Since

$$\lim_{n \to \infty} \frac{1}{1 + (x - n)^2} = 0,$$

the sequence $(f_n)_{n \geq 1}$ converges simply to the zero function $f$ given by $f(x) = 0$ for all $x \in \mathbb{R}$. However, since the the maximum of each $f_n$ is 1, we have

$$d_\infty(f_n, f) = 1 \quad \text{for all } n \geq 1,$$

so the sequence $(f_n)_{n \geq 1}$ does not converge uniformly to the zero function.

**Example 1.2.** Pick any positive real $\alpha > 0$. For each $n \geq 1$, let $f_n: \mathbb{R} \to \mathbb{R}$ be the piecewise affine function defined as follows:

$$f_n(x) = \begin{cases} 
0 & \text{if } x \leq 0 \text{ or } x \geq 1/n \\
(2n)\alpha x & \text{if } 0 \leq x \leq 1/(2n) \\
2n^{\alpha}(1 - n.x) & \text{if } 1/(2n) \leq x \leq 1/n.
\end{cases}$$

See Figure 1.7.
Figure 1.5: The colored functions $f_n(x) = n \sin \left( \frac{x}{n} \right)$, over the domain $[0, 2\pi]$, converge uniformly to the black line $f(x) = x$.

Figure 1.6: The bell curve graphs of Example 1.1; $g(x)$ in brown; $f_1(x)$ in red; $f_2(x)$ in purple; $f_3(x)$ in blue.

For every $x > 0$, there is some $n$ such that $1/n < x$, so $\lim_{n \to \infty} f_n(x) = 0$ for $x > 0$, and since $f_n(x) = 0$ for $x \leq 0$, we see that the sequence $(f_n)_{n \geq 1}$ converges simply to the zero function $f$. However, the maximum of $f_n$ is $n^\alpha$ (for $x = 1/(2n)$) so

$$d_\infty(f_n, f) = n^\alpha,$$

and $\lim_{n \to \infty} d_\infty(f_n, f) = \infty$, so the sequence $(f_n)_{n \geq 1}$ does not converge uniformly to the zero function.

Observe that convergence in the metric space of bounded functions $((B^E)_b, d_\infty)$ is the uniform convergence of sequences of functions. Similarly, convergence in the normed vector space of bounded functions $((B^E)_b, \| \|_\infty)$ is the uniform convergence of sequences of functions. For this reason, the topology on $(B^E)_b$ induced by the metric $d_\infty$ (or the norm $\| \|_\infty$) is sometimes called the topology of uniform convergence.

If $E$ is a topological space, it is useful to define the following notion of convergence.
1.2. SIMPLE AND UNIFORM CONVERGENCE

Figure 1.7: The piecewise affine functions of Example 1.2 with $\alpha = 3; f_1(x)$ in magenta; $f_2(x)$ in red; $f_3(x)$ in purple; $f_3(x)$ in blue. Each $f_n(x)$ has a symmetrical triangular peak. As $n$ increases, the peak becomes taller and thinner.

**Definition 1.6.** Let $E$ be a topological space and let $(F,d)$ be a metric space. A sequence $(f_n)_{n \geq 1}$ of functions $f_n: E \to F$ converges locally uniformly to a function $f: E \to F$ if for every $x \in E$, there is some open subset $U$ of $E$ containing $x$ such that for every $\epsilon > 0$, there is some $N > 0$ such that

$$d(f_n(x), f(x)) < \epsilon \quad \text{for all } n \geq N \text{ and for all } x \in U.$$ 

If $E$ is locally compact, it is easy to see that a sequence $(f_n)_{n \geq 1}$ converges locally uniformly iff it converges uniformly on every compact subset of $E$.

If a sequence $(f_n)_{n \geq 1}$ of continuous functions converges simply to a function $f$, the limit $f$ is not necessarily continuous. For example, the functions $f_n: [0, 1] \to \mathbb{R}$ given by $f_n(x) = x^n$ are continuous, and the sequence $(f_n)_{n \geq 1}$ converges simply to the discontinuous function $f: [0, 1] \to \mathbb{R}$ given by

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

as evidenced by Figure 1.8.

The following theorem gives sufficient conditions for the limit of a sequence of continuous functions to be continuous.

**Theorem 1.2.** Let $E$ be a topological space, and $(F,d)$ be a metric space, and let $(f_n)_{n \geq 1}$ be a sequence of functions $f_n: E \to F$ converging locally uniformly to a function $f: E \to F$. Then the following properties hold:

1. If the functions $f_n$ are continuous at some point $a \in E$, then the limit $f$ is also continuous at $a$. 
CHAPTER 1. FUNCTION SPACES OFTEN ENCOUNTERED

Figure 1.8: The sequence of functions $f_n(x) = x^n$ over $[0, 1]$ converges pointwise to the discontinuous green graph.

(2) If the functions $f_n$ are continuous (on the whole of $E$), then the limit $f$ is also continuous (on the whole of $E$).

(3) If $E$ is metric space, the sequence $(f_n)_{n \geq 1}$ converges uniformly to $f$, and the $f_n$ are uniformly continuous on $E$, then the limit $f$ is also uniformly continuous on $E$.

The proof of Theorem 1.2 can be found in Schwartz [38] (Chapter XV, Section 4, Theorem 1).

Here are a few applications of Theorem 1.2.

**Definition 1.7.** Let $E$ be a topological space, and let $(F, d)$ be a metric space. The metric subspace of $((F^E)_b, d_\infty)$ consisting of all continuous bounded functions $f : E \to F$ is denoted $C_b(E; F)$. If $(E, \|\|)$ is a normed vector space, the normed subspace of $((F^E)_b, \|\|_\infty)$ consisting of all continuous bounded functions $f : E \to F$ is also denoted $C_b(E; F)$.

**Proposition 1.3.** Let $E$ be a topological space, and let $(F, d)$ be a metric space. The metric subspace $C_b(E; F)$ of $((F^E)_b, d_\infty)$ is closed. If $(F, d)$ is a complete metric space, then $(C_b(E; F), d_\infty)$ is also complete.

**Proposition 1.4.** Let $E$ be a topological space, and let $(F, \|\|)$ be a normed vector space. The normed subspace $C_b(E; F)$ of $((F^E)_b, \|\|_\infty)$ is closed. If $(F, \|\|)$ is a complete normed vector space, then $(C_b(E; F), \|\|_\infty)$ is also complete.

An important special case of Proposition 1.4 is the case where $F = \mathbb{R}$ or $F = \mathbb{C}$, namely, our functions are real-valued continuous and bounded functions $f : E \to \mathbb{R}$, or complex-valued continuous and bounded functions $f : E \to \mathbb{C}$. The spaces of functions $(C_b(E; \mathbb{R}), d_\infty)$ and $(C_b(E; \mathbb{C}, \|\|_\infty)$ are complete.
If \( E \) is compact and if \((F, \|\|)\) is a complete normed vector space, then every continuous function \( f: E \to F \) is bounded. As a consequence, the space \( C(E; F) \) of continuous functions \( f: E \to F \) is complete.

A subspace of \((F^{\mathbb{R}})_b\), where \( F \) is a Banach space (a complete normed vector space), that plays an important role in the theory of the Riemann integral, is the space of regulated functions.

### 1.3 Regulated Functions

Recall that there are four kinds of intervals of \( \mathbb{R} \): \((a, b)\), \([a, b)\), \((a, b]\), and \([a, b]\), with \( a < b \). By convention, \((a, b) = [a, b]\) if \( a = \infty \), and \((a, b) = (a, b]\) if \( b = \infty \).

**Definition 1.8.** Let \( I \) be an interval of \( \mathbb{R} \), and let \( F \) be a metric space (or a normed vector space). Given a function \( f: I \to F \), for any \( x \in I \) with \( x \neq b \), we say that \( f \) has a limit to the right in \( x \) if \( \lim_{y \in I, y > x} f(y) \) exists as \( y \in I \) tends to \( x \) from above. This limit is denoted by \( f(x^+) \). For any \( x \in I \) with \( x \neq a \), we say that \( f \) has a limit to the left in \( x \) if \( \lim_{y \in I, y < x} f(y) \) exists as \( y \in I \) tends to \( x \) from below. This limit is denoted by \( f(x^-) \). Given any interval \( I \), a function \( f: I \to F \) is a regulated function (or ruled function) if it has a left limit and a right limit for every \( x \in I \). If \( F \) is a normed vector space, a function \( f: \mathbb{R} \to F \) is a regulated function (or ruled function) if there is some interval \( I \) such that \( f \) vanishes outside \( I \), and the restriction \( f: I \to F \) of \( f \) to \( I \) is regulated. See Figure 1.9.

![Figure 1.9](image)

Figure 1.9: An illustration of a regulated function \( f: \mathbb{R} \to \mathbb{R} \). This function has three discontinuities \( x_1, x_2, \) and \( x_3 \), each of the first kind. Note that \( f(x_1^-) = y_4, f(x_1^+) = f(x_1) = y_2, f(x_2^-) = f(x_2) = y_3, f(x_2^+) = y_6, f(x_3^-) = y_3, f(x_3^+) = y_5 \), yet \( f(x_3) = y_1 \).

The notion of a regulated function can also be defined in terms of certain kinds of discontinuities.
**Definition 1.9.** Let $I$ be an interval of $\mathbb{R}$, and let $F$ be a metric space (or a normed vector space). Given a function $f : I \to F$, we say that a point $x \in I$ is a *discontinuity of the first kind* if the left limit $f(x^-)$ and the right limit $f(+) both exist, but $f(x^-) \neq f(x)$ or $f(x^+) \neq f(x)$.

It is clear that a function $f : I \to F$ is regulated iff for every $x \in I$, either $f$ is continuous or $x$ is a discontinuity of the first kind. Thus every continuous function is a regulated function. It is also easy to see that a monotonic function $f : I \to \mathbb{R}$ is a regulated function.

The function $f : \mathbb{R} \to \mathbb{R}$ defined by

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

is discontinuous at $x = 0$, but this is not a discontinuity of the first kind. See Figure 1.10.

The following result is shown in Schwartz [40] (Chapter III, Section 2, Theorem 3.2.3).

**Proposition 1.5.** If $f : I \to F$ is a regulated function (where $F$ is a metric space), then $f$ has at most countably many discontinuities of the first kind.

Regulated functions on a closed and bounded interval $[a, b]$ must be bounded. As a consequence, they arise as limits of uniformly convergent sequences of step functions.

**Definition 1.10.** A function $f : \mathbb{R} \to F$ (where $F$ is any set) is a *step function* if there is a finite sequence $(a_0, a_1, \ldots, a_n)$ of reals such that $a_k < a_{k+1}$ for $k = 0, \ldots, n$, and $f$ is constant on each of the open intervals $(-\infty, a_0)$, $(a_k, a_{k+1})$ for $k = 0, \ldots, n$, and $(a_n, +\infty)$. The sequence $(a_0, a_1, \ldots, a_n)$ is called an *admissible subdivision* for $f$. See Figure 1.11.
Observe that Definition 1.10 does not make any restriction on the values \( f(a_k) \), but a step function is regulated. Also, a given step function admit infinitely many admissible subdivisions, by refining a given subdivision. By a step function \( f : [a, b] \rightarrow F \), we mean a step function such that \( f(x) = 0 \) for all \( x \leq a \) and for all \( x \geq b \).

The following result is easy to prove.

**Proposition 1.6.** If \( F \) is a vector space, then the set of step functions \( f : \mathbb{R} \rightarrow F \) is a vector space denoted by \( \text{Step}(\mathbb{R}; F) \). The set of step functions \( f : [a, b] \rightarrow F \) is also vector space denoted by \( \text{Step}([a, b]; F) \).

The following proposition is much more interesting.

**Proposition 1.7.** Let \( F \) be a metric space and let \( [a, b] \) be a closed and bounded interval. Then every regulated function \( f : [a, b] \rightarrow F \) is the limit of a uniformly convergent sequence \( (f_n)_{n \geq 1} \) of step functions \( f_n : [a, b] \rightarrow F \). Furthermore, if \( F \) is a complete metric space, then the limit of any uniformly convergent sequence \( (f_n)_{n \geq 1} \) of step functions is a regulated function.

The proof of Proposition 1.7 is given in Schwartz [40] (Chapter III, Section 2, Theorem 3.2.9).

As a corollary of Proposition 1.7, if \( F \) is a complete metric space, then the space of regulated functions on \( [a, b] \) is closed in \( (F^{[a,b]})_b \), and the space of step functions on \( [a, b] \) is dense in the space of regulated functions on \( [a, b] \). Thus if \( F \) is complete, since \( (F^{[a,b]})_b \) is complete, the space of regulated function on \( [a, b] \) is also complete.

Another corollary of Proposition 1.7 is that every continuous function \( f : [a, b] \rightarrow F \) to a metric space \( F \) is the limit of a uniformly convergent sequence \( (f_n)_{n \geq 1} \) of step functions \( f_n : [a, b] \rightarrow F \).
If $F$ is a vector space, the set of regulated functions defined on the closed and bounded interval $[a, b]$ is a vector space denoted by $\text{Reg}([a, b]; F)$. Then, Proposition 1.7 implies the following result.

**Proposition 1.8.** Let $F$ be a complete normed vector space. The space $\text{Reg}([a, b]; F)$ of regulated functions on $[a, b]$ is complete, and the space $\text{Step}([a, b]; F)$ is dense in $\text{Reg}([a, b]; F)$.

Step functions can be used to define the Riemann integral. To do so it is convenient to consider functions of finite support.

**Definition 1.11.** Given any function $f : E \to F$, where $E$ is a topological space and $F$ is a vector space, the support $\text{supp}(f)$ of $f$ is the closure of the subset of $E$ where $f$ is nonzero, that is, $\text{supp}(f) = \{ x \in E \mid f(x) \neq 0 \}$. The function $f$ has compact support if its support $\text{supp}(f)$ is compact. If $E$ is Hausdorff, this is equivalent to saying that $f$ vanishes outside some compact subset $K$ of $E$. See Figure 1.12.

![Figure 1.12: The graph of $f : \mathbb{R}^2 \to \mathbb{R}$ with compact support $\text{supp} = \overline{B(0, 2)} = \{ (x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 2 \}$.](image)

If a step function $f$ has compact support, then we assume that $f$ vanishes on $(-\infty, a_0)$ and on $(a_n, +\infty)$ for any admissible subdivision $(a_0, a_1, \ldots, a_n)$ for $f$.

It is easy to see that the set of continuous functions $f : E \to F$ with compact support is a vector space.
1.3. REGULATED FUNCTIONS

**Definition 1.12.** The vector space of continuous functions \( f : E \to F \) with compact support is denoted by \( \mathcal{C}_c(E; F) \), or \( \mathcal{K}(E; F) \). For every compact subset \( K \) of \( E \), we denote by \( \mathcal{K}(K; F) \) the space of continuous functions whose support is contained in \( K \). Then

\[
\mathcal{K}(E; F) = \bigcup_{K \subseteq E, K \text{ compact}} \mathcal{K}(K; F).
\]

Observe that every function in \( \mathcal{K}(E; F) \) is bounded, that is, \( \mathcal{K}(E; F) \subseteq \mathcal{C}_b(E; F) \).

If \( F = \mathbb{R} \) or \( F = \mathbb{C} \), then we write \( \mathcal{K}_\mathbb{R}(E) \) or \( \mathcal{K}_\mathbb{C}(E) \) for \( \mathcal{K}(E; F) \). Radon functionals are certain kinds of linear forms on \( \mathcal{K}_\mathbb{C}(\mathbb{C}) \).

If \( (F, \| \cdot \|) \) is a Banach space and \( K \) is a fixed compact subset of \( E \), then so is \( \mathcal{K}(K; F) \) (for the sup norm \( \| \cdot \|_\infty \)), because it is closed in \( \mathcal{C}_b(E; F) \). However, the normed vector space \((\mathcal{K}(E; F), \| \cdot \|_\infty)\) is not complete!

**Example 1.3.** For every \( n \geq 1 \), consider the function \( u_n : \mathbb{R} \to \mathbb{R} \) defined as follows:

\[
u_n(x) = \begin{cases} 
1 & \text{if } -n \leq x \leq n \\
x + n + 1 & \text{if } -(n + 1) \leq x \leq -n \\
-x + n + 1 & \text{if } n \leq x \leq n + 1 \\
0 & \text{if } |x| \geq n + 1.
\end{cases}
\]

Now consider the sequence of functions \((f_n)\) given by

\[f_n(x) = u_n e^{-|x|}.\]

Each function \( f_n \) is continuous and has compact support \([- (n + 1), n + 1]\), and it is easy to show that the sequence \((f_n)\) converges uniformly to the function \( f \) given by \( f(x) = e^{-|x|} \), but \( f \) does not have compact support. The problem is that the domains of the functions \( f_n \), although compact, keep growing as \( n \) goes to infinity. See Figure 1.13.

Example 1.3 shows that the normed vector space \((\mathcal{K}(E; F), \| \cdot \|_\infty)\) is not closed in the complete normed vector space \((\mathcal{C}_b(E; F), \| \cdot \|_\infty)\). It would be useful to identify the closure \( \overline{\mathcal{K}(E; F)} \) of \( \mathcal{K}(E; F) \) in \( \mathcal{C}_b(E; F) \), and this can indeed be done when \( E \) is locally compact.

Assume that \( f \) belongs to the closure \( \overline{\mathcal{K}(E; F)} \) of \( \mathcal{K}(E; F) \). This means that there is a sequence \((f_n)\) of functions \( f_n \in \mathcal{K}(E; F) \) such that \( \lim_{n \to \infty} \| f - f_n \|_\infty = 0 \), so for every \( \epsilon > 0 \), there is some \( n \geq 1 \) such that \( \| f(x) - f_n(x) \| \leq \epsilon \) for all \( x \in E \); and since \( f_n \) has compact support, there is some compact subset \( K \) of \( E \) such that \( \| f(x) \| \leq \epsilon \) for all \( x \in E - K \). This suggests the following definition.

**Definition 1.13.** The subspace of \( \mathcal{C}_0(E; F) \), denoted \( \mathcal{C}_0(E; F) \), consisting of the continuous functions \( f \) such that for every \( \epsilon > 0 \), there is some compact subset \( K \) of \( E \) such that \( \| f(x) \| \leq \epsilon \) for all \( x \in E - K \), is called the space of continuous functions which tend to 0 at infinity.
Observe that if \( X = \mathbb{R} \), then a function \( f \in \mathcal{C}_0(\mathbb{R}; F) \) does indeed have the property that \( \lim_{x \to -\infty} f(x) = \lim_{x \to +\infty} f(x) = 0 \).

We showed that \( \mathcal{K}(E; F) \subseteq \mathcal{C}_0(E; F) \). It is easy to see that \( \mathcal{C}_0(E; F) \) is closed in \( \mathcal{C}_0(E; \mathbb{C}) \), so it is complete. If \( E \) is locally compact, then we have the following result from Dieudonné [12] (Chapter XIII, Section 20) and Rudin [36] (Chapter 3, Theorem 3.17).

**Proposition 1.9.** If \( E \) is locally compact, then \( \mathcal{K}(E; \mathbb{C}) \) is dense in \( \mathcal{C}_0(E; \mathbb{C}) \).

**Proof.** Pick any \( f \) in \( \mathcal{C}_0(E; \mathbb{C}) \). For every \( \epsilon > 0 \), there is a compact subset \( K \) of \( E \) such that \( |f(x)| < \epsilon \) outside of \( K \). By Proposition 13.39, there is continuous function \( g: E \to [0, 1] \) with compact support such that \( g(x) = 1 \) for all \( x \in K \). Clearly \( fg \in \mathcal{K}_c(E) \), and \( \|fg - f\|_\infty < \epsilon \). This shows that \( \mathcal{K}(E; \mathbb{C}) \) is dense in \( \mathcal{C}_0(E; \mathbb{C}) \). 

In summary, we have the strict inclusions

\[
\mathcal{K}(E; \mathbb{C}) \subset \mathcal{C}_0(E; \mathbb{C}) \subset \mathcal{C}_b(E; \mathbb{C}),
\]

with \( \mathcal{C}_0(E; \mathbb{C}) \) and \( \mathcal{C}_b(E; \mathbb{C}) \) complete, and \( \mathcal{K}(E; \mathbb{C}) \) dense in \( \mathcal{C}_0(E; \mathbb{C}) \). It turns out that the space of continuous linear forms on \( \mathcal{C}_0(E; \mathbb{C}) \) is isomorphic to the space of bounded Radon functionals.
1.3. REGULATED FUNCTIONS

A modified version of step functions involving a measure will be used to define the integral on a measure space.
Chapter 2

The Riemann Integral

Let \( f: [a, b] \to \mathbb{R} \) be a continuous function. Intuitively, the Riemann integral \( \int_a^b f(t) dt \) is the area of the surface “under the curve” \( t \mapsto f(t) \) from \( x = a \) to \( x = b \). It can be approximated by the sum \( s_T(f) \) (called Cauchy-Riemann sum) of the areas \((t_{k+1} - t_k)f(t_k)\) of \( n \geq 1 \) narrow rectangles, where \( T = (t_0, t_1, \ldots, t_n) \) is any sequence of reals such that \( t_0 = a, t_n = b \) and \( t_k < t_{k+1} \), for \( k = 0, \ldots, n - 1 \); see Figure 2.1. The fact that the function \( f \) is continuous on the compact interval \([a, b]\) implies that the sums \( s_T(f) \) have a limit when the diameter of the subdivision tends to zero (see Definition 2.1), which means the maximum of the distances \( t_{k+1} - t_k \) tends to zero (as \( n \) goes to infinity), and this limit is independent of the subdivision. Thus we can define the Riemann integral \( \int_a^b f(t) dt \) as this common limit. The mapping \( f \mapsto \int_a^b f(t) dt \) is a positive linear form on the space of continuous functions on \([a, b]\). This procedure applies unchanged to continuous functions \( f: [a, b] \to F \), where \( F \) is a complete normed vector space.

The method for constructing the integral of a continuous function can be adapted to define the integral of regulated functions (see Definition 1.8). We proceed in two steps:

1. (1) The method of Cauchy-Riemann sums is easily adapted to define the notion of integral for a step function (see Definition 1.10). This yields a mapping \( \int: \text{Step}([a, b]; F) \to F \) which is easily seen to be linear and continuous.

2. (2) By Proposition 1.8, the vector space \( \text{Step}([a, b]; F) \) of steps functions over \([a, b]\) is dense in \( \text{Reg}([a, b]; F) \), the space of regulated functions over \([a, b]\), and \( \text{Reg}([a, b]; F) \) is complete. By Theorem 13.69, the continuous linear map \( \int: \text{Step}([a, b]; F) \to F \) has a unique extension \( \int: \text{Reg}([a, b]; F) \to F \) to \( \text{Reg}([a, b]; F) \), which is also continuous and linear. This is how the integral of a regulated function is defined.

In summary, we define an “obvious” notion of integral on the simple set \( \text{Step}([a, b]; F) \). It is a linear and continuous mapping, so we extend it by continuity to the bigger space \( \text{Reg}([a, b]; F) \) in which \( \text{Step}([a, b]; F) \) is dense.
2.1 Riemann Integral of a Continuous Function

Definition 2.1. Let \( a < b \) be any two reals. A set \( T = \{t_0, t_1, \ldots, t_n\} \) of reals such that \( t_0 = a, \ t_n = b \) and \( t_k < t_{k+1} \), for \( k = 0, \ldots, n - 1 \), is called a subdivision of \([a, b]\). The diameter \( \delta(T) \) of \( T \) is defined by

\[ \delta(T) = \max_{0 \leq k \leq n-1} (t_{k+1} - t_k). \]

Given a continuous function \( f : [a, b] \to \mathbb{R} \), define the Cauchy–Riemann sum \( s_T(f) \) by

\[ s_T(f) = \sum_{k=0}^{n-1} (t_{k+1} - t_k) f(t_k). \]

See Figure 2.1

![Figure 2.1: The Cauchy–Riemann sum \( s_T(f) = \sum_{k=0}^{n-1} (t_{k+1} - t_k) f(t_k) \) is the area represented by the pastel shaded boxes.](image)

Observe that

\[ \sum_{k=0}^{n-1} (t_{k+1} - t_k) = b - a. \]

We immediately check that \( s_T \) is a linear form on the set of continuous functions on \([a, b]\). Furthermore, if \( f \geq 0 \), which means that \( f(t) \geq 0 \) for all \( t \in [a, b] \), then \( s_T(f) \geq 0 \).

The question is, as the subdivision \( T \) becomes finer and finer, in the sense that \( \delta(T) \) becomes smaller and smaller (which means that \( n \) gets bigger and bigger), do the sums \( s_T(f) \) have a limit?

The answer is yes.
2.1. RIEMANN INTEGRAL OF A CONTINUOUS FUNCTION

The reason is that a continuous function on a compact interval \([a, b]\) is uniformly continuous, and this implies that for any sequence \((T_m)\) of subdivisions such that \(\delta(T_m) \to 0\) as \(m\) goes to infinity, the sums \(s_{T_m}(f)\) form a Cauchy sequence, as we now explain.

**Proposition 2.1.** Let \(f : [a, b] \to \mathbb{R}\) be a continuous function defined on a closed and bounded (compact) interval \([a, b]\). For every \(\epsilon > 0\), there is some \(\eta > 0\) such that for any two subdivisions \(T\) and \(T'\) of \([a, b]\) such that \(\delta(T) < \eta\) and \(\delta(T') < \eta\), we have

\[
|s_T(f) - s_{T'}(f)| < \epsilon.
\]

**Proof.** Since a continuous function on \([a, b]\) is actually uniformly continuous, for any \(\epsilon > 0\), we can find some \(\eta > 0\) such that

\[
|f(x) - f(x')| < \epsilon/(2(b-a)) \quad \text{for all } x, x' \in [a, b] \text{ such that } |x - x'| < \eta.
\]

If \(T = \{t_0, t_1, \ldots, t_n\}\) and \(T' = \{t'_0, t'_1, \ldots, t'_{n'}\}\), let \(T'' = T \cup T'\) and let \(T''_k = T'' \cap [t_k, t_{k+1}]\), more precisely, \(T''_k = \{s_0, s_1, \ldots, s_r\}\), with \(s_0 = t_k\), \(s_r = t_{k+1}\), and \(r = 0\) if the above set on the right-hand side is empty, for \(k = 0, \ldots, n - 1\).

Then we immediately check that

\[
T'' = \bigcup_{k=0}^{n-1} T''_k, \quad \text{and} \quad s_{T''}(f) = \sum_{k=0}^{n-1} s_{T''_k}(f).
\]

See Figure 2.2.

---

![Figure 2.2](image-url)

Figure 2.2: An illustration of the refinement \(s_{T''}(f)\) utilized in the proof of Proposition 2.1. Note that \(T\) is given by the black dots while \(T'\) is given by the brown dots.
Since \( s_{T_k''}(f) \) is of the form
\[
s_{T_k''}(f) = \sum_{i=0}^{r-1} (s_{i+1} - s_i) f(s_i),
\]
where \( t_k \leq s_i \leq t_{k+1} \) for \( i = 0, \ldots, r \), and since \( \sum_{i=0}^{r-1} (s_{i+1} - s_i) = s_r - s_0 = t_{k+1} - t_k \), we have
\[
|s_{T_k''}(f) - (t_{k+1} - t_k) f(t_k)| = \left| \sum_{i=0}^{r-1} (s_{i+1} - s_i) f(s_i) - (t_{k+1} - t_k) f(t_k) \right|
= \left| \sum_{i=0}^{r-1} (s_{i+1} - s_i) (f(s_i) - f(t_k)) \right|
\leq \sum_{i=0}^{r-1} |(s_{i+1} - s_i)||f(s_i) - f(t_k)|
< \sum_{i=0}^{r-1} |(s_{i+1} - s_i)| \frac{\epsilon}{2(b - a)}
= (t_{k+1} - t_k) \frac{\epsilon}{2(b - a)}.
\]
As a consequence, we obtain
\[
|s_T(f) - s_{T''}(f)| = \left| \sum_{k=0}^{n-1} (t_{k+1} - t_k) f(t_k) - \sum_{k=0}^{n-1} s_{T_k''}(f) \right|
\leq \sum_{k=0}^{n-1} |s_{T_k''}(f) - (t_{k+1} - t_k) f(t_k)|
< \sum_{k=0}^{n-1} (t_{k+1} - t_k) \frac{\epsilon}{2(b - a)}
\leq \frac{\epsilon}{2},
\]
that is,
\[
|s_T(f) - s_{T''}(f)| < \frac{\epsilon}{2}.
\]
By a similar argument applied to \( T' \), we obtain
\[
|s_{T'}(f) - s_{T''}(f)| < \frac{\epsilon}{2}.
\]
But then we obtain
\[
|s_T(f) - s_{T'}(f)| = |s_T(f) - s_{T''}(f) + s_{T''}(f) - s_{T'}(f)|
\leq |s_T(f) - s_{T''}(f)| + |s_{T''}(f) - s_{T'}(f)|
< \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,
\]
as claimed.

**Remark:** It is easy to check that the proof of Proposition 2.1 is still valid if we use more general Cauchy-Riemann sums. Namely, given a subdivision $T = \{t_0, t_1, \ldots, t_n\}$ of $[a, b]$, and any choice of reals $\theta_1, \ldots, \theta_n$ such that $t_k \leq \theta_{k+1} \leq t_{k+1}$ for $k = 0, \ldots, n-1$, define $s_T(f)$ as

$$s_T(f) = \sum_{k=0}^{n-1} (t_{k+1} - t_k) f(\theta_k).$$

Proposition 2.1 implies the following result, which establishes the existence of the Riemann integral of a continuous function defined on a closed and bounded (compact) interval $[a, b]$.

**Theorem 2.2.** Let $f: [a, b] \to \mathbb{R}$ be a continuous function defined on a closed and bounded (compact) interval $[a, b]$. For every sequence $\mathcal{T} = (T_m)_{m \geq 1}$ of subdivisions of $[a, b]$ such that $\lim_{m \to \infty} \delta(T_m) = 0$, the sequence $(s_{T_m}(f))_{m \geq 1}$ is a Cauchy sequence, and thus has a limit $S_T(f)$. For any two sequences $\mathcal{T} = (T_m)_{m \geq 1}$ and $\mathcal{T}' = (T'_m)_{m \geq 1}$ of subdivisions of $[a, b]$, if $\lim_{m \to \infty} \delta(T_m) = 0$ and $\lim_{m \to \infty} \delta(T'_m) = 0$, then $S_T(f) = S_{T'}(f)$, that is, the limit of the sequence $(s_{T_m}(f))$ is independent of the sequence $\mathcal{T} = (T_m)_{m \geq 1}$ such that $\lim_{m \to \infty} \delta(T_m) = 0$.

**Proof.** Pick any $\epsilon > 0$, an let $\eta > 0$ be some number given by Proposition 2.1, such that for any two subdivisions $T$ and $T'$ of $[a, b]$ such that $\delta(T) < \eta$ and $\delta(T') < \eta$, we have

$$|s_T(f) - s_{T'}(f)| < \epsilon.$$

Since $\lim_{m \to \infty} \delta(T_m) = 0$, there is some $N > 0$ such that for all $m, n \geq N$, we have $\delta(T_m) < \eta$ and $\delta(T_n) < \eta$, which by the definition of $\eta$, implies that

$$|s_{T_m}(f) - s_{T_n}(f)| < \epsilon \quad \text{for all} \ m, n \geq N.$$

Therefore, $(s_{T_m}(f)))$ is a Cauchy sequence. Since $\mathbb{R}$ is a complete metric space, this sequence has a limit $S_T(f)$.

Since by hypothesis $\lim_{m \to \infty} \delta(T_m) = 0$ and $\lim_{m \to \infty} \delta(T'_m) = 0$, there is some $N > 0$ such that for all $m \geq N$, we have $\delta(T_m) < \eta$ and $\delta(T'_m) < \eta$, so by Proposition 2.1,

$$|s_{T_m}(f) - s_{T'_m}(f)| < \epsilon \quad \text{for all} \ m \geq N,$$

which shows that the Cauchy sequences $(s_{T_m}(f))$ and $(s_{T'_m}(f))$ have the same limit.

**Theorem 2.2** also holds for the more general Cauchy-Riemann sums defined in the Remark after Proposition 2.1.

Theorem 2.2 justifies the following definition.
Definition 2.2. Let $f: [a, b] \to \mathbb{R}$ be a continuous function defined on a closed and bounded (compact) interval $[a, b]$. The common limit $S_T(f)$ of the Cauchy sequences $(s_{T_m}(f))_{m \geq 1}$, for all sequences $T = (T_m)_{m \geq 1}$ of subdivisions of $[a, b]$ such that $\lim_{m \to \infty} \delta(T_m) = 0$, is called the Riemann integral of $f$, and is denoted by $\int_a^b f(t)\,dt$.

The following are basic properties of the Riemann integral, which are easy to prove (using suitable subdivisions of $[a, b]$):

1. The mapping $f \mapsto \int_a^b f(t)\,dt$ is a linear form on the space of continuous functions on $[a, b]$. This means that for any two continuous functions $f, g: [a, b] \to \mathbb{R}$ and any scalar $\lambda \in \mathbb{R}$,

$$
\int_a^b (f + g)(t)\,dt = \int_a^b f(t)\,dt + \int_a^b g(t)\,dt
$$

$$
\int_a^b (\lambda f)(t)\,dt = \lambda \int_a^b f(t)\,dt,
$$

where, as usual, $f + g$ is the function given by $(f + g)(t) = f(t) + g(t)$, and $\lambda f$ is the function given by $(\lambda f)(t) = \lambda f(t)$, for all $t \in [a, b]$. Furthermore, it is a positive linear form, which means that if $f \geq 0$, then $\int_a^b f(t)\,dt \geq 0$. These seemingly innocuous properties turn out to be very important. Indeed, we will see later how the notion of integral on a locally compact space can be defined in terms of such linear forms (Radon functionals).

2. $\left| \int_a^b f(t)\,dt \right| \leq \int_a^b |f(t)|\,dt \leq (b - a) \max_{t \in [a, b]} |f(t)|$.

3. If $f \geq 0$ and $f(t) > 0$ for some $t \in [a, b]$, then $\int_a^b f(t)\,dt > 0$.

4. If $a < b < c$, then

$$
\int_a^b f(t)\,dt + \int_b^c f(t)\,dt = \int_a^c f(t)\,dt.
$$

5. If $H: [a, b] \to \mathbb{R}$ is the function given by

$$
H(x) = \int_a^x f(t)\,dt,
$$

then $H$ is differentiable on $[a, b]$ and $H'(x) = f(x)$ (the so-called first fundamental theorem of calculus).

The process that we just described only requires that the codomain be complete and that a continuous function $f: [a, b] \to F$ be uniformly continuous. We also need the linear
combinations $\sum_{k=0}^{n-1} (t_{k+1} - t_k) f(t_k)$ to make sense, so $F$ should be a vector space. If we assume that $F$ is a complete normed vector space (a Banach space), then the Riemann integral of a vector-valued function $f : [a, b] \to F$ can be defined by using the method that we just presented.

In the next section, we show how to define the integral of function with discontinuities, provided that these discontinuities are “reasonable.” For this, a new crucial idea is needed: to define the integral on a class of simple functions with a finite number of reasonable discontinuities, and then to extend the integral to a bigger class of functions by taking limits of simple functions. For this process to work, the bigger space of functions should be complete.

2.2 The Riemann Integral of Regulated Functions

In this section we show how to define the integral of regulated functions $f : [a, b] \to F$, where $F$ is any complete normed vector space, in particular $\mathbb{R}$ or any finite-dimensional vector space (real or complex).

The first key ingredient is that the method of Cauchy-Riemann sums can be immediately adapted to define the notion of integral for a step function. The mapping $\int : \text{Step}([a, b]; F) \to F$ is easily seen to be linear and continuous.

The second key ingredient is that, by Proposition 1.8, the vector space $\text{Step}([a, b]; F)$ of steps functions over $[a, b]$ is dense in $\text{Reg}([a, b]; F)$, the space of regulated functions over $[a, b]$, and $\text{Reg}([a, b]; F)$ is complete, where $[a, b]$ is a closed and bounded interval.

Then, because $\text{Step}([a, b]; F)$ is dense in $\text{Reg}([a, b]; F)$, and $\text{Reg}([a, b]; F)$ is complete, by Theorem 13.69, the continuous linear map $\int : \text{Step}([a, b]; F) \to F$ has a unique extension $\int : \text{Reg}([a, b]; F) \to F$ to $\text{Reg}([a, b]; F)$, which is also continuous and linear. This is how the integral of a regulated function is defined.

Thus it remains to define the integral of a step function.

**Definition 2.3.** Let $f : [a, b] \to F$ be a step function. For any admissible subdivision $T = (a_0, a_1, \ldots, a_n)$ for $f$, for any sequence $\xi = (\xi_1, \ldots, \xi_n)$ of reals such that $\xi_{k+1} \in (a_k, a_{k+1})$ for $k = 0, \ldots, n - 1$, define $s_{T, \xi}(f)$ by

$$s_{T,\xi}(f) = \sum_{k=0}^{n-1} (a_{k+1} - a_k) f(\xi_{k+1}).$$

See Figure 2.3.

The above is a linear combination of vectors in $F$, and since $F$ is a vector space, it is well defined. Note that because $\xi_{k+1} \in (a_k, a_{k+1})$, $s_{T,\xi}(f)$ does not depend on the value of $f$ at the $a_k$. For simplicity of language, we refer to a pair $(T, \xi)$ as in Definition 2.3 as an admissible pair for $f$. 
CHAPTER 2. THE RIEMANN INTEGRAL

Figure 2.3: An illustration of $s_{T,\xi}(f) = \sum_{k=0}^{3}(a_{k+1} - a_k)f(\xi_{k+1})$ for step function $f: [a, b] \to \mathbb{R}$.

The problem with the above definition of $s_{T,\xi}(f)$ is that it depends on the admissible subdivision $T$, and on $\xi$, but because $f$ is a step function, it is constant on each interval $(a_k, a_{k+1})$, so in fact $s_{T,\xi}(f)$ is independent of the admissible pair $(T, \xi)$.

Proposition 2.3. Given a step function $f: [a, b] \to F$, for any two admissible pairs $(T, \xi)$ and $(T', \xi')$ for $f$, we have $s_{T,\xi}(f) = s_{T',\xi'}(f)$.

Proposition 2.3 is proved by using an admissible pair which is finer than both $(T, \xi)$ and $(T', \xi')$. The details are left to the reader, or see Schwartz [41] (Chapter V, Section §1).

Proposition 2.3 justifies the following definition.

Definition 2.4. Let $f: [a, b] \to F$ be a step function. The integral of $f$, denoted $\int_{[a,b]} f$, is the common value of the sum $s_{T,\xi}(f)$, for any any admissible pair $(T, \xi)$ for $f$.

The following proposition follows almost immediately from the definitions.

Proposition 2.4. The map $\int : \text{Step}([a,b]; F) \to F$, where $\int f = \int_{[a,b]} f$ is the integral defined in Definition 2.4, is linear. Furthermore, we have

$$\left\| \int_{[a,b]} f \right\| \leq \int_{[a,b]} \|f\| \quad \text{and} \quad \left\| \int_{[a,b]} f \right\| \leq (b-a) \|f\|_{\infty},$$

where $\|f\|$ means the real-valued function $x \mapsto \|f(x)\|$. If $f = \mathbb{R}$ and and if $f \geq 0$, then $\int_{[a,b]} f \geq 0$.

Proposition 2.4 shows that the map $\int : \text{Step}([a,b]; F) \to F$ is linear and continuous. As we explained earlier, by Theorem 13.69, the map $\int : \text{Step}([a,b]; F) \to F$ has a unique extension $\int : \text{Reg}([a,b]; F) \to F$ to $\text{Reg}([a,b]; F)$, which is also linear and continuous.
Definition 2.5. The integral \( \int_{[a,b]} f \) of any regulated function \( f \in \text{Reg}([a,b]; F) \) is equal to \( \int f \), where \( \int : \text{Reg}([a,b]; F) \to F \) is the unique linear and continuous extension of the linear and continuous map \( \int : \text{Step}([a,b]; F) \to F \). This integral is called the \textit{Riemann integral} of the regulated function \( f \).

Definition 2.5 is not very constructive. It turns out that the Riemann integral of a regulated function can be defined more directly in terms of generalized Riemann sums. This approach is presented in Schwartz [41] (Chapter V, Section §1).

Note that we actually haven’t defined the notion of Riemann-integrable function. What we did is to exhibit a family of functions, the regulated functions, which are Riemann-integrable function. The notion of Riemann-integrable function is defined in various books, including Schwartz [41]. This can be done using the notion of upper integral \( \int^* f \), which is defined for a positive function \( f \in \mathcal{K}(\mathbb{R}, F) \) as the infimum of the integrals of the step functions that bound \( f \) from above.

The space of Riemann-integrable functions contains other functions besides the regulated functions. For example, functions with compact support which are continuous except at finitely many points, are Riemann-integrable. The function \( x \mapsto \sin(1/x) \) is such a function (with value 0 at \( x = 0 \)). It is Riemann-integrable on \([0, 1]\), even though 0 is not a discontinuity of the first kind.

The method of this section, which consists in defining the notion of integral for a “big” set of functions, such as \( \text{Reg}([a,b]; F) \), by first defining a notion of integral on a very simple set of functions for which the definition is obvious, such as \( \text{Step}([a,b]; F) \), and then to extend the integral on \( \text{Step}([a,b]; F) \) to a notion of integral on \( \text{Reg}([a,b]; F) \) using a completion process, is a key idea. In this situation, we are lucky that \( \text{Reg}([a,b]; F) \) is complete.

In order to define a notion of integral for functions defined on a domain \( X \) which is more general than a compact interval \([a,b]\) of \( \mathbb{R} \), we can proceed as above, but some additional structure on \( X \) is needed to define step functions and the notion of integral of step functions. This new ingredient is the notion of \textit{measure}. The other technical difficulty is that the completion of the space of generalized step functions is not a space identifiable with a space of familiar functions. By Theorem 13.68, the completion always exists, but its elements are equivalence classes of functions, so it will take some work to exhibit this space as a set of functions.
Chapter 3

Measure Theory; Basic Notions

Let $X$ be a nonempty set. Intuitively, a measure on $X$ is a function $\mu$ that assigns a nonnegative real number $\mu(A)$ to every subset $A$ in some specified nonempty collection $\mathcal{A}$ of subsets of $X$, where $\mu(A)$ is a generalization of the notion of length, area, or volume. For example, any natural measure $\mu$ on $\mathbb{R}$ should have the property that $\mu((a,b)) = \mu([a,b]) = b - a$ for all $a \leq b$. It is natural to require that if a subset $A$ is sliced into countably many pairwise disjoint small pieces $A_i$, then $\mu(A) = \mu\left(\bigcup_{n=1}^{\infty} A_i\right) = \sum_{n=1}^{\infty} \mu(A_i)$. This property is called $\sigma$-additivity. Then the family $\mathcal{A}$ of subsets on which $\mu$ is defined should be closed under countable unions. It is also natural to require $\mathcal{A}$ to be closed under complementation. This leads to the important notion of a $\sigma$-algebra, which is closed under complementation and countable unions. The weaker notion in which only closure under complementation and closure under finite union is that of an algebra. In general, it is not easy to construct nontrivial $\sigma$-algebras, so it is useful to have tools to do so. A pair $(X, \mathcal{A})$ consisting of a nonempty set and a $\sigma$-algebra $\mathcal{A}$, is called a measurable space.

Given any nonempty family $S$ of subsets of $X$, there is a smallest $\sigma$-algebra $\mathcal{A}(S)$ containing $S$. If $X$ is a topological space, then the $\sigma$-algebra $\mathcal{B}(X)$ containing the open subsets of $X$ is an important $\sigma$-algebra called the Borel $\sigma$-algebra.

The notion of monotone class is also useful to construct $\sigma$-algebras. Given any nonempty family $S$ of subsets of $X$, there is a smallest monotone class $\mathcal{M}(S)$ containing $S$. Given an algebra $\mathcal{B}$, the smallest $\sigma$-algebra $\mathcal{A}(\mathcal{B})$ containing $\mathcal{B}$ and the smallest monotone class $\mathcal{M}(\mathcal{B})$ containing $\mathcal{B}$ are identical: $\mathcal{A}(\mathcal{B}) = \mathcal{M}(\mathcal{B})$.

Next we define (positive) measures on a $\sigma$-algebra. A triple $(X, \mathcal{A}, \mu)$ consisting of a nonempty set, a $\sigma$-algebra $\mathcal{A}$, and a measure $\mu$ on $\mathcal{A}$, is called a measure space. We investigate a few properties of measures. In particular, we show that every measure can be extended to a complete measure, which means that all $A \in \mathcal{A}$, if $\mu(A) = 0$, then $B \in \mathcal{A}$ for all $B \subseteq A$.

As we said earlier, it is not easy to construct nontrivial measures. A very useful concept to achieve this is the notion of outer measure, introduced in Section 3.3. Outer measures are defined for all subsets of $X$, which makes them much easier to construct. In particular, we construct the Lebesgue outer measure $\mu^*_L$. 

33
A fundamental theorem due to Carathéodory shows that every outer measure induces a measure space; see Theorem 3.11.

By applying Theorem 3.11 to the outer measure \( \mu^*_L \) we obtain the \( \sigma \)-algebra \( \mathcal{L}(\mathbb{R}) \) of Lebesgue measurable sets and the Lebesgue measure \( \mu_L \); see Section 3.4. The Borel \( \sigma \)-algebra \( \mathcal{B}(\mathbb{R}) \) is properly contained in the \( \sigma \)-algebra \( \mathcal{L}(\mathbb{R}) \) of Lebesgue measurable sets, and there are subsets of \( \mathbb{R} \) that are not Lebesgue measurable sets (assuming the axiom of choice). We also discuss various regularity properties of the Lebesgue measure.

### 3.1 \( \sigma \)-Algebras, Measures

Let \( X \) be a nonempty set. We would like to define the notion of “measure” for the subsets of \( X \), in such a way that familiar properties of the notion of length, area, or volume, of polyhedral objects in \( \mathbb{R}, \mathbb{R}^2 \) or \( \mathbb{R}^3 \), hold. The measure \( m(A) \) of a subset of \( X \) should be nonnegative, but we have to allow “big” objects to have infinite measure so it is desirable to extend the nonnegative real numbers by adding a new element corresponding to infinity.

Technically, we define \( \mathbb{R}_+ \) as the union

\[
\mathbb{R}_+ = \{ \alpha \in \mathbb{R} \mid \alpha \geq 0 \} \cup \{ +\infty \} = \mathbb{R}_+ \cup \{ +\infty \},
\]

where \( +\infty \) is not in \( \mathbb{R}_+ \), and we assume that the following properties hold:

1. \( \alpha < +\infty \), for all \( \alpha \in \mathbb{R}_+ \),
2. \( \alpha + (+\infty) = (+\infty) + \alpha = +\infty \), for all \( \alpha \in \mathbb{R}_+ \),
3. \( \alpha \cdot (+\infty) = (+\infty) \cdot \alpha = +\infty \), for all \( \alpha \in \mathbb{R}_+ \),
4. \( 0 \cdot (+\infty) = (+\infty) \cdot 0 = 0 \),
5. If \( (\alpha_i)_{i \geq n} \) is a sequence with \( \alpha_i \in \mathbb{R}_+ \), and if \( \alpha_i = +\infty \) for some \( i \), then \( \sum_{i=1}^{\infty} \alpha_i = +\infty \).

The set \( \mathbb{R}_+ \) is also denoted by \([0, +\infty]\).

In this chapter, we closely follow Halmos [22]. Other nice (but concise) presentations can be found in Rudin [36] and Lang [24]. A very detailed presentation is given in Schwartz [41].

An “ideal measure” should be a function \( m \) satisfying the following properties:

1. \( m : 2^X \to [0, +\infty] \), that is, \( m \) is defined for all subsets of \( X \).
2. \( m(\emptyset) = 0 \).
3. For any countable sequence \( (A_i)_{i \geq 1} \) of subsets \( A_i \) of \( X \) such that \( A_i \cap A_j = \emptyset \) for all \( i \neq j \),

\[
m \left( \bigcup_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} m(A_i).
\]

This property is called \( \sigma \)-additivity.
The intuition behind \( \sigma \)-additivity is that if we slice an object \( A \) into countably many pairwise disjoint small pieces \( A_i \), then the measure \( m(A) \) of \( A \) should be the sum of the measures \( m(A_i) \) of the pieces \( A_i \).

Observe that by choosing a sequence \((A_i)_{i \geq 1}\) such that \( A_j = \emptyset \) for all \( j > n \), and \( A_i \cap A_j = \emptyset \) if \( i \neq j \), we obtain the property

\[
m \left( \bigcup_{i=1}^{n} A_i \right) = \sum_{i=1}^{n} m(A_i),
\]

known as additivity.

For any two subsets \( A \) and \( B \), if \( A \subseteq B \), we can write \( B = A \cup (B - A) \), with \( A \cap (B - A) = \emptyset \), so by additivity,

\[
m(B) = m(A) + m(B - A),
\]

and since \( m(B - A) \geq 0 \), we obtain

\[
m(A) \leq m(B).
\]

We claim that the following property holds.

**Proposition 3.1.** If a function \( m \) satisfies Properties (1–3) above, then for any countable sequence \((A_i)_{i \geq 1}\) of (not necessarily pairwise disjoint) subsets \( A_i \) of \( X \),

\[
m \left( \bigcup_{i=1}^{\infty} A_i \right) \leq \sum_{i=1}^{\infty} m(A_i).
\]

**Proof.** Define the sequence \((B_i)\) of subsets of \( X \) as follows: \( B_1 = A_1 \), \( B_2 = A_2 - A_1 \),..., \( B_i = A_i - \left( \bigcup_{j=1}^{i-1} A_j \right) \), for all \( i \geq 2 \). See Figure 3.1.

![Figure 3.1](image-url)
It is easy to check that $\bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} A_i$, $B_i \cap B_j = \emptyset$ for all $i \neq j$, and $m(B_i) \leq m(A_i)$ for all $i \geq 1$, so by $\sigma$-additivity,

$$m\left(\bigcup_{i=1}^{\infty} A_i\right) = m\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} m(B_i) \leq \sum_{i=1}^{\infty} m(A_i),$$

as claimed.

In general, for an arbitrary set $X$, there may be no function $m$ satisfying Properties (1–3) for all subsets of $X$, as well as certain desirable properties. For example, there is no such function on $2^\mathbb{R}$ such that $m([a, b]) = b - a$ for every interval $[a, b]$.

Thus we are led to relax some of these conditions. There are two options:

1. The first option is to relax (3) by replacing it by the result of Proposition 3.1, namely

   $$(3') \text{ For any countable sequence } (A_i)_{i \geq 1} \text{ of subsets } A_i \text{ of } X,$$

   $$m\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} m(A_i).$$

   This approach leads to outer measures, and is discussed in Section 3.3.

2. Condition (3) is highly desirable, so the second option is to restrict the domain of $m$ to be a proper family of subsets of $X$; the right notion is that of a $\sigma$-algebra.

The notion of a $\sigma$-algebra is more important than the notion of outer measure, which is needed for technical reasons. Thus we now consider Option 2, and define $\sigma$-algebras. Once the notion of $\sigma$-algebra is defined, we will be able to define the abstract notion of a measure (see Definition 3.9), which is the crucial ingredient in defining a general notion of integral.

**Definition 3.1.** Let $X$ be any nonempty set. A family $\mathcal{A}$ of subsets of $X$ is a $\sigma$-algebra if it satisfies the following conditions:

(A1) $X \in \mathcal{A}$.

(A2) For every subset $A$ of $X$, if $A \in \mathcal{A}$, then $X - A \in \mathcal{A}$ (closure under complementation).

($\sigma$-A3) For every countable family $(A_i)_{i \geq 1}$ of subsets of $X$, if $A_i \in \mathcal{A}$ for all $i \geq 1$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$ (closure under countable unions).

From (A1) and (A2), we see that $\emptyset \in \mathcal{A}$. From (A2) and ($\sigma$-A3) and the fact that $A = X - (X - A)$ and $\bigcap_{i=1}^{\infty} A_i = X - (\bigcup_{i=1}^{\infty} (X - A_i))$, if $A_i \in \mathcal{A}$ for all $i \geq 1$, then $\bigcap_{i=1}^{\infty} A_i \in \mathcal{A}$ (closure under countable intersections). In particular, if we let $A_i = \emptyset$ for all $i \geq 3$, we see that if $A_1 \in \mathcal{A}$ and $A_2 \in \mathcal{A}$, then $A_1 \cup A_2 \in \mathcal{A}$ and $A_1 \cap A_2 \in \mathcal{A}$. Since $A_1 - A_2 = A_1 \cap (X - A_2)$, we also have $A_1 - A_2 \in \mathcal{A}$.

Axiom ($\sigma$-A3) is a strong condition, and this is the reason why it is not easy to construct nontrivial $\sigma$-algebras. There are two extreme $\sigma$-algebras:
3.1. \(\sigma\)-ALGEBRAS, MEASURES

1. \(\mathcal{A} = \{\emptyset, X\}\).

2. \(\mathcal{A} = 2^X\), the family of all subsets of \(X\).

Interesting \(\sigma\)-algebra lie in-between.

Remarks:

1. Some authors use the term \(\sigma\)-field instead of \(\sigma\)-algebra. This is a rather unfortunate terminology, because in algebra, a field is a set with two operations that have identity elements. Here the operations are union and intersection, but there is no identity element for intersection.

2. If we weaken Condition \(\sigma\)-A3 to finite unions, then we obtain a structure called an algebra (or boolean algebra of sets).

Definition 3.2. Let \(X\) be any nonempty set. A family \(\mathcal{A}\) of subsets of \(X\) is an algebra (or boolean algebra of sets) if it satisfies the following conditions:

(A1) \(X \in \mathcal{A}\).

(A2) For every subset \(A\) of \(X\), if \(A \in \mathcal{A}\), then \(X - A \in \mathcal{A}\) (closure under complementation).

(A3) For every finite family \((A_i)_{i=1}^n\) of subsets of \(X\), if \(A_i \in \mathcal{A}\) for all \(i = 1, \ldots, n\), then \(\bigcup_{i=1}^n A_i \in \mathcal{A}\) (closure under finite unions).

As in the case of \(\sigma\)-algebras, an algebra contains \(\emptyset\) and is closed under (finite) union and intersection. In the construction of a product of measurable spaces, another notion of algebra will come up. These are the semi-algebras.

Definition 3.3. Let \(X\) be any nonempty set. A family \(\mathcal{S}\) of subsets of \(X\) is a semi-algebra if it satisfies the following conditions:

(S1) \(X, \emptyset \in \mathcal{S}\).

(S2) For all \(A, B \in \mathcal{S}\), we have \(A \cap B \in \mathcal{S}\).

(S3) For all \(A \in \mathcal{S}\), we have \(X - A = X_1 \cup \cdots \cup X_n\), for finitely many pairwise disjoint subsets \(X_i \in \mathcal{S}\).

Example 3.1. First consider the family of intervals of \(\mathbb{R}\) of the form \([a, b)\), with \(a \leq b\), where \(a = -\infty\) or \(b = \infty\) is allowed. By convention, let \([a, b) = \emptyset\) if \(a > b\). This is a semi-algebra, because

\([a_1, b_1) \cap [a_2, b_2) = [\max(a_1, a_2), \min(b_1, b_2))\),

and

\(X - [a, b) = (-\infty, a) \cup [b, \infty)\).
Next, let $X$ and $Y$ be two nonempty sets, and let $\mathcal{A}$ be an algebra on $X$ and let $\mathcal{B}$ be an algebra on $Y$. Define the set $\mathcal{R}$ of rectangles in $X \times Y$ as follows:

$$\mathcal{R} = \{ A \times B \in X \times Y \mid A \in \mathcal{A}, B \in \mathcal{B} \}.$$ 

It is easy to check that $\mathcal{R}$ is a semi-algebra. For example,

$$(A_1 \times B_1) \cap (A_2 \times B_2) = (A_1 \cap A_2) \times (B_1 \cap B_2),$$

and

$$(X \times Y) - (A \times B) = ((X - A) \times (Y - B)) \cup ((X - A) \times B) \cup (A \times (Y - B)).$$

Then it can be shown that the set $\mathcal{B}(\mathcal{R})$ of finite unions of pairwise disjoint sets in $\mathcal{R}$ is the smallest algebra containing the semi-algebra $\mathcal{R}$. This algebra will be used to construct the product of measurable spaces.

The following result can be shown.

**Proposition 3.2.** Given a semi-algebra $\mathcal{S}$, the smallest algebra $\mathcal{B}(\mathcal{S})$ containing $\mathcal{S}$ is the family of finite unions of pairwise disjoint subsets in $\mathcal{S}$.

**Definition 3.4.** Let $X$ be any nonempty set. A pair $(X, \mathcal{A})$ where $\mathcal{A}$ is a $\sigma$-algebra of subsets of $X$ is called a measurable space. The subsets of $X$ that belong to $\mathcal{A}$ are called the measurable subsets of $X$.

**Proposition 3.3.** Let $X$ be any nonempty set, and let $\mathcal{S}$ be any nonempty family of subsets of $X$. Then there is a $\sigma$-algebra $\mathcal{A}(\mathcal{S})$ with the following properties:

(a) $\mathcal{S} \subseteq \mathcal{A}(\mathcal{S})$.

(b) If $\mathcal{A}'$ is any $\sigma$-algebra such that $\mathcal{S} \subseteq \mathcal{A}'$, then $\mathcal{A}(\mathcal{S}) \subseteq \mathcal{A}'$.

This means that $\mathcal{A}(\mathcal{S})$ is the smallest $\sigma$-algebra containing $\mathcal{S}$.

**Definition 3.5.** Let $X$ be any nonempty set, and let $\mathcal{S}$ be any nonempty family of subsets of $X$. The smallest $\sigma$-algebra $\mathcal{A}(\mathcal{S})$ containing $\mathcal{S}$ is called the $\sigma$-algebra generated by $\mathcal{S}$.

The $\sigma$-algebra $\mathcal{A}(\mathcal{S})$ is the intersection of the family of all $\sigma$-algebras containing $\mathcal{S}$. This family is nonempty since $2^X$ belongs to it. This way of defining $\mathcal{A}(\mathcal{S})$ is highly nonconstructive. A bottom-up construction of $\mathcal{A}(\mathcal{S})$ can be performed, but to guarantee closure under infinite unions, transfinite induction is required; see Schwartz [41] (Chapter V, Section §2).

An important example arises when $X$ is a topological space $(X, \mathcal{O})$.

**Definition 3.6.** Let $(X, \mathcal{O})$ be a topological space. The $\sigma$-algebra $\mathcal{B}(X)$ generated by the family $\mathcal{O}$ of open sets is called the Borel $\sigma$-algebra of $X$. The subsets in $\mathcal{B}(X)$ are called Borel sets.
All open subsets and all closed sets are Borel sets. Countably infinite unions of closed
sets and countable infinite intersections of open sets are Borel sets. But there are many more
Borel sets.

Another way to construct \(\sigma\)-algebras is to use algebras and monotone classes.

**Definition 3.7.** Let \( X \) be any nonempty set. A nonempty family \( \mathcal{M} \) of subsets of \( X \) is a
**monotone class** if for every countable family \( (A_i)_{i \geq 1} \) of subsets of \( X \), if \( A_i \in \mathcal{M} \) for all \( i \geq 1 \) then:

1. If \( A_i \subseteq A_{i+1} \) for all \( i \geq 1 \), then \( \bigcup_{i=1}^{\infty} A_i \in \mathcal{M} \). See Figure 3.2, Diagram (i).
2. If \( A_{i+1} \subseteq A_i \) for all \( i \geq 1 \), then \( \bigcap_{i=1}^{\infty} A_i \in \mathcal{M} \). See Figure 3.2, Diagram (ii).

Figure 3.2: The rose colored sets of Figure (i) satisfy the increasing nesting condition of
\( A_i \subseteq A_{i+1} \), while the periwinkle sets of Figure (ii) satisfy the decreasing nesting condition
\( A_{i+1} \subseteq A_i \).

**Proposition 3.4.** Let \( X \) be any nonempty set. For any algebra \( \mathcal{B} \), if \( \mathcal{B} \) is a monotone class,
then \( \mathcal{B} \) is a \( \sigma\)-algebra.

**Proof.** Let \( (A_i)_{i \geq 1} \) be a countable family of subsets of \( X \), such that \( A_i \in \mathcal{B} \) for all \( i \geq 1 \). Since
\( \mathcal{B} \) is an algebra, it is closed under finite unions, so \( B_i = \bigcup_{i=1}^{n} A_i \in \mathcal{B} \), and obviously \( B_i \subseteq B_{i+1} \)
for all \( i \geq 1 \), and \( \bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i \). Since \( \mathcal{B} \) is a monotone class, \( \bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i \in \mathcal{B} \). \( \square \)

Here is a version of Proposition 3.3 for monotone classes.

**Proposition 3.5.** Let \( X \) be any nonempty set, and let \( S \) be any nonempty family of subsets
of \( X \). Then there is a monotone class \( \mathcal{M}(S) \) with the following properties:

(a) \( S \subseteq \mathcal{M}(S) \).
(b) If $\mathcal{M}$ is any monotone class such that $\mathcal{S} \subseteq \mathcal{M}$, then $\mathcal{M}(\mathcal{S}) \subseteq \mathcal{M}$.

This means that $\mathcal{M}(\mathcal{S})$ is the smallest monotone class containing $\mathcal{S}$.

**Definition 3.8.** Let $X$ be any nonempty set, and let $\mathcal{S}$ be any nonempty family of subsets of $X$. The smallest monotone class $\mathcal{M}(\mathcal{S})$ containing $\mathcal{S}$ is called the **monotone class generated by $\mathcal{S}$**.

The following theorem yields another way of generating a $\sigma$-algebra from an algebra.

**Theorem 3.6.** Let $X$ be any nonempty set. For any algebra $\mathcal{B}$, the $\sigma$-algebra $\mathcal{A}(\mathcal{B})$ generated by $\mathcal{B}$ and the monotone class $\mathcal{M}(\mathcal{B})$ generated by $\mathcal{B}$ are identical; that is,

$$\mathcal{A}(\mathcal{B}) = \mathcal{M}(\mathcal{B}).$$

We now come to the very important notion of measure.

**Definition 3.9.** Let $X$ be any nonempty set. A **measure** on $X$ is a map $\mu$ satisfying the following properties:

1. $\mu: \mathcal{A} \rightarrow [0, +\infty]$, where $\mathcal{A}$ is a $\sigma$-algebra of subsets of $X$.
2. $\mu(\emptyset) = 0$.
3. For any countable sequence $(A_i)_{i \geq 1}$ of subsets $A_i$ of $\mathcal{A}$ such that $A_i \cap A_j = \emptyset$ for all $i \neq j$,

$$\mu \left( \bigcup_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} \mu(A_i).$$

This property is called $\sigma$-additivity.

A **measure space** is a triple $(X, \mathcal{A}, \mu)$, where $(X, \mathcal{A})$ is a measurable space and $\mu$ is a measure on $X$.

**Remarks:**

1. The degenerate situation where $\mu(A) = +\infty$ for all nonempty subsets in $\mathcal{A}$ is allowed. If $\mu$ is nontrivial, which means that $\mathcal{A}$ possesses some nonempty subset $A$ such that $\mu(A)$ is finite, then by letting $A_1 = A$ and $A_i = \emptyset$ for all $i \geq 2$, by $(\mu3)$ we get $\mu(A) = \mu(A) + \sum_{i=2}^{\infty} \mu(\emptyset)$, which implies $\mu(\emptyset) = 0$. In this situation Axiom $(\mu2)$ is unnecessary. Rudin makes the assumption that a measure is nontrivial; see [36].

2. A measure $\mu$ is also called a **positive measure**, to stress that its range is nonnegative. There are more general measures taking their values in $\mathbb{R}$ or $\mathbb{C}$, even a in Banach space. For such measures, Condition $(\mu3)$ needs to be slightly strengthened.
3. Some authors used the term measured space instead of measure space!

**Definition 3.10.** Let \((X, \mathcal{A}, \mu)\) be a measure space. The measure \(\mu\) is finite if \(\mu(X)\) is finite. If \(\mu : \mathcal{A} \to [0, 1]\) and if \(\mu(X) = 1\), then \((X, \mathcal{A}, \mu)\) is called a probability space. The measure \(\mu\) is \(\sigma\)-finite if there exist a countable family \((A_i)_{i \geq 1}\) of subsets \(A_i \in \mathcal{A}\) such that \(X = \bigcup_{i=1}^{\infty} A_i\) and \(\mu(A_i)\) is finite for all \(i \geq 1\). The measure \(\mu\) is complete if for all \(A \in \mathcal{A}\), if \(\mu(A) = 0\), then \(B \in \mathcal{A}\) for all \(B \subseteq A\). A subset \(A \in \mathcal{A}\) such that \(\mu(A) = 0\) is called a set of measure zero.

**Example 3.2.** Let \(X\) be any nonempty set, and consider the \(\sigma\)-algebra \(\mathcal{A} = 2^X\). the map \(\mu : 2^X \to [0, +\infty]\) given by

\[
\mu(A) = \begin{cases} 
|A| & \text{if } A \text{ is finite} \\ +\infty & \text{if } A \text{ is infinite}
\end{cases}
\]

is a measure called the counting measure on \(X\).

Here is a summary of useful properties of measures.

**Proposition 3.7.** Let \((X, \mathcal{A}, \mu)\) be a measure space. The following properties hold:

1. For any finite sequence \((A_1, \ldots, A_n)\) of subsets \(A_i \in \mathcal{A}\) such that \(A_i \cap A_j = \emptyset\) whenever \(i \neq j\), we have

\[
\mu \left( \bigcup_{i=1}^{n} A_i \right) = \sum_{i=1}^{n} \mu(A_i).
\]

2. For any two subsets \(A, B\) of \(X\), if \(A, B \in \mathcal{A}\) and if \(A \subseteq B\), then \(\mu(A) \leq \mu(B)\).

3. For any countable sequence \((A_i)_{i \geq 1}\) of subsets \(A_i \in \mathcal{A}\),

\[
\mu \left( \bigcup_{i=1}^{\infty} A_i \right) \leq \sum_{i=1}^{\infty} \mu(A_i).
\]

4. For any countable sequence \((A_i)_{i \geq 1}\) of subsets \(A_i \in \mathcal{A}\), if \(A_{i+1} \subseteq A_i\) for all \(i \geq 1\) and if \(\mu(A_1)\) is finite, then

\[
\mu \left( \bigcap_{i=1}^{n} A_i \right) = \lim_{n \to \infty} \mu(A_i).
\]

5. For any countable sequence \((A_i)_{i \geq 1}\) of subsets \(A_i \in \mathcal{A}\), if \(A_i \subseteq A_{i+1}\) for all \(i \geq 1\), then

\[
\mu \left( \bigcup_{i=1}^{n} A_i \right) = \lim_{n \to \infty} \mu(A_i).
\]
**CHAPTER 3. MEASURE THEORY; BASIC NOTIONS**

**Proof.** The proof of (1) and (2) is identical to the proof given just before Proposition 3.1, and (3) is Proposition 3.1. We prove (4), leaving the proof of (5) as an exercise.

We can write

\[ A_n = \left( \bigcap_{i=1}^{\infty} A_i \right) \cup \left( \bigcup_{i=n}^{\infty} (A_i - A_{i+1}) \right), \]

a union of pairwise disjoint subsets since \( A_{i+1} \subseteq A_i \) for all \( i \geq 1 \). By \((\mu_3)\), we have

\[ \mu \left( \bigcap_{i=1}^{\infty} A_i \right) + \sum_{i=n}^{\infty} \mu(A_i - A_{i+1}) = \mu(A_n) \leq \mu(A_1) < +\infty, \]

since \( A_{i+1} \subseteq A_i \) for all \( i \geq 1 \) and since \( \mu(A_1) \) is assumed to be finite. See Figure 3.3.

![Figure 3.3: Decomposing the decreasing nested sequences of periwinkle sets into disjoint rings. Note \( A_1 = (A_1 - A_2) \cup (A_2 - A_3) \cup \cdots \cup (A_n - A_{n+1}) \cup A_{n+1} \).](image)

Consequently, for \( n = 1 \) we deduce that the series \( \sum_{i=1}^{\infty} \mu(A_i - A_{i+1}) \) converges, which implies that

\[ \lim_{n \to \infty} \sum_{i=n}^{\infty} \mu(A_i - A_{i+1}) = 0. \]

Since

\[ \mu \left( \bigcap_{i=1}^{\infty} A_i \right) + \sum_{i=n}^{\infty} \mu(A_i - A_{i+1}) = \mu(A_n), \]

we conclude that \( \mu \left( \bigcap_{i=1}^{n} A_i \right) = \lim_{n \to \infty} \mu(A_i). \)

The following result shows that every measure can be completed; this is technically useful.

**Proposition 3.8.** Let \((X, \mathcal{A}, \mu)\) be a measure space. A measure space \((X, \overline{\mathcal{A}}, \overline{\mu})\) with the following properties can be constructed:
3.2. NULL SUBSETS AND PROPERTIES HOLDING ALMOST EVERYWHERE

(a) $A \subseteq \overline{A}$.

(b) $\overline{\mu}$ extends $\mu$; that is, $\overline{\mu}(A) = \mu(A)$ for all $A \in \mathcal{A}$.

(c) The measure $\overline{\mu}$ is complete.

We force the completeness property by defining $\overline{\mathcal{A}}$ as follows:

$$\overline{\mathcal{A}} = \{ \overline{A} \subseteq X \mid (\exists A, A' \in \mathcal{A})(\exists B \subseteq A')(\overline{A} = A \cup B, \mu(A') = 0) \}.$$ 

The measure $\overline{\mu}$ is defined such that

$$\overline{\mu}(\overline{A}) = \overline{\mu}(A \cup B) = \mu(A).$$

See Figure 3.4.

Figure 3.4: A schematic illustration of a set in $\overline{\mathcal{A}}$. The magenta set $A$ has positive measure, while the grayish set $A'$, and all of its subsets, including $B$, have zero measure. Then $\overline{\mathcal{A}} = A \cup B$.

The verification that $\overline{\mathcal{A}}$ is a $\sigma$-algebra with the required properties and that $\overline{\mu}$ is a measure with the required properties is tedious (among other things, one need to check that $\overline{\mu}(\overline{A})$ does not depend on the representation of $\overline{A}$).

The measure $\overline{\mu}$ given by Proposition 3.8 is called the completed measure of $\mu$.

3.2 Null Subsets and Properties Holding Almost Everywhere

One of the secrets of measure theory is that subsets of measure zero should be ignored. Since a measure is not necessarily complete the correct technical definition is as follows.
Definition 3.11. Let \((X, \mathcal{A}, \mu)\) be a measure space. A subset \(E \subseteq X\) is null\(^1\) (or negligible) if there is some \(A \in \mathcal{A}\) such that \(E \subseteq A\) and \(\mu(A) = 0\). A property \(P\) of the elements of \(X\) holds almost everywhere, abbreviated holds a.e., if the subset were it fails is null; that is, the set \(\{x \in X \mid P(x) = \text{false}\}\) is null.

To be very precise, we should say \(\mu\)-null and holds \(\mu\)-a.e., since these notions depend on the measure \(\mu\). In most cases, there is no risk of confusion and we drop \(\mu\).

Observe that if the measure \(\mu\) is complete, then a subset \(E \subseteq X\) is null iff \(\mu(E) = 0\), and a property \(P\) holds a.e. iff \(\mu(\{x \in X \mid P(x) = \text{false}\}) = 0\). In general, a null set may either be measurable or nonmeasurable, and a nonmeasurable has no reason to be null, but may be null.

Here are a few properties of null sets.

Proposition 3.9. Let \((X, \mathcal{A}, \mu)\) be a measure space. Every subset of a null set is null. Every countable union of null sets is a null set.

Proof. The first property follows immediately from the definition. Let \((A_i)_{i \geq 1}\) be a countable family of null sets. There are subsets \(B_i \in \mathcal{A}\) such that \(A_i \subseteq B_i\) and \(\mu(B_i) = 0\) for all \(i \geq 1\). We have
\[
\bigcup_{i=1}^{\infty} A_i \subseteq \bigcup_{i=1}^{\infty} B_i \in \mathcal{A}
\]
because \(\mathcal{A}\) is a \(\sigma\)-algebra, so it remains to show that \(\bigcup_{i=1}^{\infty} B_i\) has measure zero. For this, observe that
\[
0 \leq \mu\left(\bigcup_{i=1}^{\infty} B_i\right) \leq \sum_{i=1}^{\infty} \mu(B_i) = 0,
\]
so \(\mu(\bigcup_{i=1}^{\infty} B_i) = 0\), as desired. \(\square\)

Let \(P\) and \(P'\) be two properties of \(X\). If \(P\) implies \(P'\) and if \(P\) holds a.e., then \(P'\) holds a.e.

Definition 3.12. Consider the set of functions \(f : X \to \mathbb{R}\), where \((X, \mathcal{A}, \mu)\) is a measure space. We say that \(f\) and \(g\) are equal a.e. if the set \(\{x \in X \mid f(x) \neq g(x)\}\) is null. Write \(f = g\) (a.e.).

It is an easy exercise to show that equality a.e. is an equivalence relation.

It should be observed that the notion of equality a.e. is more subtle than one might think.

\(^1\)Beware that in measure theory, the notion of null set has more than one meaning. Some authors mean something different from what we define here.
3.3. CONSTRUCTION OF A MEASURE FROM AN OUTER MEASURE

Example 3.3. For example, consider the function \( \chi_Q : \mathbb{R} \to \mathbb{R} \), given by

\[
\chi_Q(x) = \begin{cases} 
1 & \text{if } x \in \mathbb{Q} \\
0 & \text{if } x \notin \mathbb{Q}.
\end{cases}
\]

In other words, \( \chi_Q \) is the characteristic function of the rationals. It is easy to see that \( \chi_Q \) is discontinuous at every point \( x \in \mathbb{R} \) (if \( x \) is irrational, then every small interval containing \( x \) contains some rational number; similarly, if \( x \) is rational, then every small interval containing \( x \) contains some irrational number, say of the form \( x + \frac{\sqrt{2}}{2n} \) for \( n \) large enough). Now, the Lebesgue measure \( \mu_L \) discussed in Section 3.4 has the property that every countable set has measure zero, so in particular \( \mathbb{Q} \) has Lebesgue measure zero. It follows that \( \chi_Q \) is equal to the zero function on a.e., and the zero function is a “very nice” function; it is infinitely differentiable.

This is the beauty of equality a.e. Given a “very bad” function, we can ignore its bad behavior on a set of measure zero, as least from the point of view of integration.

An interesting variation of \( \chi_Q \) is the following function \( D_Q : \mathbb{R} \to \mathbb{R} \), given by

\[
D_Q(x) = \begin{cases} 
\frac{1}{p} & \text{if } x = p/q \in \mathbb{Q}, q \neq 0, \gcd(p,q) = 1 \\
0 & \text{if } x \notin \mathbb{Q}.
\end{cases}
\]

It is easy to show that \( D_Q \) is discontinuous at every rational point \( x \), but is continuous at every irrational point \( x \). In fact, \( D_Q \) is a regulated function. Again \( D_Q \) is equal to the zero function a.e. (with respect to the Lebesgue measure).

A property that will play an important role is pointwise convergence a.e.

**Definition 3.13.** Let \((X, \mathcal{A}, \mu)\) be a measure space, and let \( F \) be any topological space (in most cases a normed vector space). A sequence \((f_n)_{n \geq 1}\) of functions \( f_n : X \to F \) converges pointwise a.e. to a function \( f : X \to F \) if there is a null set \( Z \subseteq X \) such that the sequence \((f_n(x))_{n \geq 1}\) converges to \( f(x) \) for all \( x \in X - Z \). See Figure 3.5.

### 3.3 Construction of a Measure from an Outer Measure

It turns out that defining explicitly a function \( m \) satisfying Conditions (2) and (3) from the beginning of Section 3.1 on a \( \sigma \)-algebra is not easy, but defining a function \( \mu^* \) on \( 2^X \) satisfying (1), (2), and (3'), is quite easy. Furthermore, given such a function \( \mu^* \), called an outer measure, there is a way of generating a \( \sigma \)-algebra and a measure on it.

If \( X \) is a locally compact topological space, then there is a way to construct a \( \sigma \)-algebra and a function \( m \) satisfying (2) and (3) on this \( \sigma \)-algebra using Radon functionals. This method will be explored in Chapter 5.

We now consider Option 1 from Section 3.1 and define outer measures.
Figure 3.5: A schematic illustration of pointwise convergence a. e. Let $X$ be the solid black line, $F = \mathbb{R}$, and $Z = Z_1 \cup Z_2 \cup Z_3$, where each $Z_i$ has measure zero. The sequence $(f_n)$ converges pointwise to the graph $f$ (in red) for all $x \in X - Z$. As shown in the bottom right corner, the magenta graph $f_{n+1}$ is "close" to $f$ outside of $Z$.

**Definition 3.14.** Given a nonempty set $X$, an *outer measure* $\mu^*$ on $X$ is a function satisfying the following properties:

$(\mu^*1)$ $\mu^*: 2^X \to [0, +\infty]$, that is, $\mu^*$ is defined for all subsets of $X$.

$(\mu^*2)$ $\mu^*(\emptyset) = 0$.

$(\mu^*3)$ For any countable sequence $(A_i)_{i \geq 1}$ of subsets $A_i$ of $X$,

$$
\mu^* \left( \bigcup_{i=1}^{\infty} A_i \right) \leq \sum_{i=1}^{\infty} \mu^*(A_i).
$$

This property is called *σ-subadditivity*.

$(\mu^*4)$ If $A \subseteq B$, then $\mu^*(A) \leq \mu^*(B)$. 
Example 3.4. (Outer measure of Dirac) Let $X$ be any nonempty set, and let $a$ be any point chosen in $X$. The map $\mu^*_a: 2^X \to [0, +\infty]$ given by
\[
\mu^*_a(A) = \begin{cases} 
1 & \text{if } a \in A \\
0 & \text{if } a \notin A 
\end{cases}
\]
is an outer measure called the outer measure of Dirac.

Here is a simple way to construct outer measures.

Proposition 3.10. Let $X$ be a nonempty set, and $\mathcal{J} \subseteq 2^X$ be a family of subsets with the following properties:

(a) $\emptyset \in \mathcal{J}$.

(b) For every subset $A$ of $X$, there is a finite sequence $(I_1, \ldots, I_n)$ of subsets $I_j \in \mathcal{J}$ such that $A \subseteq \bigcup_{j=1}^n I_j$.

Moreover, let $\lambda: \mathcal{J} \to [0, +\infty]$ be any function such that

(c) $\lambda(\emptyset) = 0$.

Then the map $\mu^*$ given by
\[
\mu^*(A) = \inf \left\{ \sum_{n=1}^{\infty} \lambda(I_n) \mid A \subseteq \bigcup_{n=1}^{\infty} I_n, I_n \in \mathcal{J} \right\}
\]
is an outer measure on $X$.

As an application of Proposition 3.10, we obtain the outer Lebesgue measure.

Example 3.5. Let $\mathcal{J}$ consist of the set of all open intervals $(a, b)$, where $a = -\infty$ or $b = +\infty$ is allowed. It is easy to see that Properties (a) and (b) of Proposition 3.10 are satisfied. Let $\lambda: \mathcal{J} \to [0, +\infty]$ be given by $\lambda((a, b)) = b - a$. Obviously, Property (c) holds. The outer measure given by Proposition 3.10 is called the outer Lebesgue measure $\mu^*_L$ on $\mathbb{R}$.

A similar construction could be performed on $\mathbb{R}^n$ by using products of open intervals $(a_1, b_1) \times \cdots \times (a_n, b_n)$ and $\lambda((a_1, b_1) \times \cdots \times (a_n, b_n)) = \prod_{i=1}^n (b_i - a_i)$.

We now state a fundamental theorem due to C. Carathéodory which gives a method for constructing a measure space from an outer measure.

Theorem 3.11. (Carathéodory) Let $\mu^*: 2^X \to [0, +\infty]$ be an outer measure. Define the family $\mathcal{A}$ of subsets of $X$ as follows:

$\mathcal{A} = \{ A \in 2^X \mid \mu^*(E) = \mu^*(E \cap A) + \mu^*(E \cap (X - A)), \text{ for all } E \subseteq X \}$.

See Figure 3.6. Then the following properties hold:
CHAPTER 3. MEASURE THEORY; BASIC NOTIONS

Figure 3.6: A schematic illustration of the Carathéodory construction of $\mathcal{A}$. The $\sigma$-algebra $\mathcal{A}$ consists of those magenta sets $A$ which “cut” (with respect to $\mu^*$) arbitrary subsets $E$ in a “nice” manner.

(a) $\mathcal{A}$ is a $\sigma$-algebra which contains all subsets $A \subseteq X$ such that $\mu^*(A) = 0$.

(b) The restriction $\mu$ of $\mu^*$ to $\mathcal{A}$ is a measure. Furthermore, $\mu$ is a complete measure.

The proof of Theorem 3.11 is not really difficult but quite long and involves some tedious verifications.

Example 3.6. If we apply Theorem 3.11 to the Dirac outer measure $\mu^*_a$ of Example 3.4, we find easily that $\mathcal{A} = 2^X$ and that $\mu = \mu^*_a$. The Dirac measure $\mu^*_a$ is ususally denoted by $\delta_a$.

If we apply Theorem 3.11 to the Lebesgue outer measure of Example 3.5, we obtain the Lebesgue measure on $\mathbb{R}$. It can be shown that the $\sigma$-algebra of Lebesgue-measurable sets obtained from the construction contains the $\sigma$-algebra of Borel sets of $\mathbb{R}$. This example is considered in slightly more details in the next section.

3.4 The Lebesgue Measure on $\mathbb{R}$

Recall that in Example 3.5 we defined the outer Lebesgue measure $\mu^*_L$ on $\mathbb{R}$. For this, we considered the set $\mathcal{I}$ consisting of all open intervals $(a, b)$, where $a = -\infty$ or $b = +\infty$ is allowed. By Proposition 3.10 applied to the function $\lambda: \mathcal{I} \to [0, +\infty]$ given by $\lambda((a, b)) =$
b − a, we obtained the outer Lebesgue measure \( \mu^*_L \) given by

\[
\mu^*_L(A) = \inf \left\{ \sum_{i=1}^{\infty} \lambda(I_n) \mid A \subseteq \bigcup_{n=1}^{\infty} I_n, \ I_n \in \mathcal{I} \right\}.
\]

By applying Theorem 3.11 to the outer measure \( \mu^*_L \), we obtain the σ-algebra \( \mathcal{L}(\mathbb{R}) \) of Lebesgue-measurable sets, and the Lebesgue measure \( \mu_L \).

The construction used by Theorem 3.11 yields very little explicit information regarding what the Lebesgue-measurable sets look like, but it is possible to describe some of them. In particular, if \( \mathcal{B}(\mathbb{R}) \) denotes the Borel σ-algebra generated by the open sets of \( \mathbb{R} \), it turns out that \( \mathcal{B}(\mathbb{R}) \subset \mathcal{L}(\mathbb{R}) \), a proper inclusion. Actually, every open subset of \( \mathbb{R} \) can be expressed as a countable disjoint union of open intervals, so the Borel σ-algebra is generated by the open intervals \((a, b)\). The following proposition gives convenient characterizations of the Borel sets.

**Proposition 3.12.** The σ-algebra \( \mathcal{B}(\mathbb{R}) \) of Borel sets of \( \mathbb{R} \) is generated by the following intervals:

1. \([a, b]\), with \( a \leq b \) finite.
2. \((a, b]\), with \( a \leq b \) finite.
3. \((a, \infty]\), with \( a \) finite.

**Proof.** (1) We know that the open intervals \((a, b)\) generate \( \mathcal{B}(\mathbb{R}) \). We have

\[
(a, \infty) = \bigcup_{n=1}^{\infty} (a, a + n), \quad (a, a) = \bigcup_{n=1}^{\infty} (a - n, a).
\]

Observe that

\[
[a, b] = (-\infty, a] \cap (b, \infty) = \bigcap_{n=1}^{\infty} \left[ a + \frac{1}{n}, b - \frac{1}{n} \right],
\]

so \([a, b] \in \mathcal{B}(\mathbb{R})\). We also have

\[
(a, b) = \bigcup_{n=1}^{\infty} \left[ a + \frac{1}{n}, b - \frac{1}{n} \right],
\]

so the closed intervals \([a, b]\) generate \( \mathcal{B}(\mathbb{R}) \).

(2) We have

\[
[a, b] = \bigcup_{n=1}^{\infty} \left[ a, b - \frac{1}{n} \right],
\]

so \([a, b] \in \mathcal{B}(\mathbb{R})\). Then

\[
(a, b) = \bigcup_{n=1}^{\infty} \left[ a + \frac{1}{n}, b \right],
\]
so the intervals \([a, b)\) generate \(\mathcal{B}(\mathbb{R})\).

(3) We already know from (1) that \((a, \infty) \in \mathcal{B}(\mathbb{R})\). This implies that 
\((-\infty, a] = (a, \infty) \in \mathcal{B}(\mathbb{R}),\)
so
\((-\infty, a) = \bigcup_{n=1}^{\infty} \left(-\infty, a - \frac{1}{n}\right) \in \mathcal{B}(\mathbb{R}),\)
and thus
\((a, b) = (-\infty, b) \cap (a, \infty),\)
so the intervals \((a, \infty)\) generate \(\mathcal{B}(\mathbb{R})\). \qed

Let use the notation \(\mathring{a, b}\) to denote any of the four types of intervals \((a, b), [a, b), (a, b], [a, b]\) (with \(a = -\infty\) or \(b = +\infty\) allowed, and \(a = b\) allowed). The following result can be shown.

**Proposition 3.13.** Let \(\mathcal{B}(\mathbb{R})\) be the Borel \(\sigma\)-algebra of open sets, \(\mathcal{L}(\mathbb{R})\) the \(\sigma\)-algebra of Lebesgue-measurable sets, and \(\mu_L\) be the Lebesgue measure, for \(\mathbb{R}\).

1. \(\mathcal{L}(\mathbb{R}) \neq 2^\mathbb{R}\); that is, there exist non-measurable sets. The proof requires the axiom of choice.

2. \(\mathcal{B}(\mathbb{R}) \subset \mathcal{L}(\mathbb{R})\), the inclusion being strict. This is because \(|\mathcal{B}(\mathbb{R})| = \aleph_1 = 2^{\aleph_0}\), but \(|\mathcal{L}(\mathbb{R})| = \aleph_2 = 2^{\aleph_1}\).

3. The Borel \(\sigma\)-algebra contains all four types of intervals, and
\[\mu_L(\mathring{a, b}) = \begin{cases} b - a & \text{if } a \neq -\infty \text{ and } b \neq +\infty \\ +\infty & \text{if } a = -\infty \text{ or } b = +\infty. \end{cases}\]

4. The restriction of the Lebesgue measure \(\mu_L\) to the Borel \(\sigma\)-algebra is a measure \(\mu_B\). The completion of the measure space \((\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_B)\) given by Proposition 3.8 gives back the measure space \((\mathbb{R}, \mathcal{L}(\mathbb{R}), \mu_L)\) of Lebesgue-measurable sets.

It is surprising how much work it takes to prove Part (3) of Proposition 3.13. As a corollary, every one-point set \(\{a\}\) has Lebesgue measure 0, and thus every countable subset has Lebesgue measure 0. There are also uncountable subsets of Lebesgue measure 0.

The Lebesgue measure also has the following regularity properties which show that every Lebesgue-measurable set can be approximated either by an open set or by a closed set.

**Proposition 3.14.** For any subset \(A\) of \(\mathbb{R}\), we have
\[\mu^*_L(A) = \inf\{\mu_L(O) \mid A \subseteq O, O \text{ is open}\}.$

For every Lebesgue-measurable set \(A \in \mathcal{L}(\mathbb{R})\), the following facts hold:
3.4. THE LEBESGUE MEASURE ON \( \mathbb{R} \)

(a) For every \( \epsilon > 0 \), there is some open subset \( O \) such that \( A \subseteq O \) and \( \mu_L(O - A) < \epsilon \).

(a) For every \( \epsilon > 0 \), there is some closed subset \( F \) such that \( F \subseteq A \) and \( \mu_L(A - F) < \epsilon \).

As a corollary of Proposition 3.14 we have the following facts: For every Lebesgue-measurable set \( A \in \mathcal{L}(\mathbb{R}) \):

(a’) \( \mu_L(A) = \inf \{ \mu_L(O) \mid A \subseteq O, \ O \text{ is open} \} \).

(b’) \( \mu_L(A) = \sup \{ \mu_L(F) \mid F \subseteq A, \ F \text{ is closed} \} \).

See Figure 3.7

![Figure 3.7](image)

Figure 3.7: A Lebesgue-measurable set \( A \) of \( \mathbb{R} \) is approximated from the “outside” by an open set \( O \); it is also approximated from the “inside” by a closed set \( F \).

It should be noted that Properties (a’) and (b’) are weaker than Properties (a) and (b), because they imply Properties (a) and (b) only when \( \mu(A) \) is finite.

It can also be shown that for every Lebesgue-measurable set \( A \in \mathcal{L}(\mathbb{R}) \), we have

\[
\mu_L(A) = \sup \{ \mu_L(K) \mid K \subseteq A, \ K \text{ is compact} \}.
\]

Proposition 3.14 also holds for the Lebesgue-measurable subsets of \( \mathbb{R}^n \).

We conclude by mentioning that if \( X \) is a topological space, given a function \( \mu \) defined on the open subsets and the compact subsets of \( X \), we can define the following maps for every subset \( A \) of \( X \):

\[
\mu^*(A) = \inf \{ \mu(O) \mid A \subseteq O, \ A \text{ is open} \}
\]

\[
\mu_*(A) = \sup \{ \mu(K) \mid K \subseteq A, \ K \text{ is compact} \}.
\]
Then the measurable subsets are those subsets $A$ of $X$ such that

$$
\mu^*(A) = \mu_*(A).
$$

It can be shown that these subsets form a $\sigma$-algebra $\mathcal{A}$, and that the map $\mu$ with domain $\mathcal{A}$ given by $\mu(A) = \mu^*(A) = \mu_*(A)$ is a measure. This is the approach using Radon measures.
Chapter 4
Integration

Given a measure space \((X, \mathcal{A}, \mu)\), we would like to define the integral of a real-valued function \(f: X \to \mathbb{R}\), or more generally of a complex-valued function \(f: X \to \mathbb{C}\), or even of a function \(f: X \to F\), where \(F\) is a normed vector space. The key idea is that the integral of a very simple function \(f\), such as a function taking only a finite number of nonzero values \(y_1, \ldots, y_n\), should be “obvious.” Namely, if \(A_i = f^{-1}(y_i)\) is the subset of \(X\) over which \(f\) has the value \(y_i\), then each \(A_i\) should be measurable (that is, \(A_i \in \mathcal{A}\)), and \(A_i\) should have finite measure, so that the expression

\[
\sum_{i=1}^{n} y_i \mu(A_i) \in F
\]

makes sense. Then we define the integral \(\int f \, d\mu\) of our simple function \(f\) as

\[
\int f \, d\mu = \sum_{i=1}^{n} \mu(A_i) y_i. \tag{\ast}
\]

Observe that the function \(f\) can be written as

\[
f = \sum_{i=1}^{n} y_i \chi_{A_i},
\]

where \(\chi_{A_i}\) is the characteristic function of the subset \(A_i\). Such a function is called a \(\mu\)-step function.

Observe that \((\ast)\) is a generalization of the notion of area under the curve. If the subsets \(A_i\) are closed adjacent intervals, then we are back to the notion of Riemann integral. However, in our new setting, the subsets \(A_i\) can be very complicated, but as long as they are measurable and have finite measure, the integral \((\ast)\) makes sense.

If we define \(\|f\|\) as

\[
\|f\| = \sum_{i=1}^{n} \|y_i\| \chi_{A_i},
\]

\[53\]
If we define $N(f)$ (remember that our set $F$ of values is a normed vector space), then the integral of $\|f\|$ is

$$\int \|f\| \, d\mu = \sum_{i=1}^{n} \mu(A_i) \|y_i\| \in \mathbb{R}_+.$$

If we define $N_1(f) = \int \|f\| \, d\mu$, then $N_1$ satisfies all the properties of a norm, except that $N_1(f) = 0$ does not necessarily imply that $f = 0$. However, $N_1(f) = 0$ iff $f = 0$ almost everywhere. The set $\text{Step}_\mu(X, \mathcal{A}, F)$ of $\mu$-step functions is a vector space, and $N_1$ is almost a norm on it; it is a semi-norm. The integral given by $(\star)$ is a linear continuous map on $\text{Step}_\mu(X, \mathcal{A}, F)$. However, the space $\text{Step}_\mu(X, \mathcal{A}, F)$ is not Cauchy-complete under the semi-norm $N_1$ (there are Cauchy sequences with respect to $N_1$ that do not have a limit). The problem then is to complete the space $\text{Step}_\mu(X, \mathcal{A}, \mu)$ and to extend the integral $(\star)$ to this bigger set of functions.

There are several ways to proceed.

1. If we let $SN$ be subspace of $\text{Step}_\mu(X, \mathcal{A}, F)$ consisting of the $\mu$-step functions equal to 0 a.e., then the quotient space $\text{Step}_\mu(X, \mathcal{A}, F) = \text{Step}_\mu(X, \mathcal{A}, \mu)/SN$ is a vector space and $N_1$ induces a (true) norm on it. Therefore we can apply the general completion theorem (Theorem 13.68) to obtain a complete normed vector space $(L_\mu(X, \mathcal{A}, F), \| \|_1)$. Since integration is a linear continuous map on $\text{Step}_\mu(X, \mathcal{A}, F)$, it extends uniquely to a linear continuous map on $L_\mu(X, \mathcal{A}, F)$.

In theory, we have achieved our goal of defining a complete normed vector space of functions containing the $\mu$-step functions for which every function is integrable. However, the completion $(L_\mu(X, \mathcal{A}, F), \| \|_1)$ is a very complicated object. It consists of equivalence classes of Cauchy sequences of functions in the quotient space $\text{Step}_\mu(X, \mathcal{A}, F)$. It would be much more convenient if the objects in $L_\mu(X, \mathcal{A}, F)$ could be described as functions, and this is indeed possible.

2. The second approach is to first define a set $\mathcal{L}_\mu(X, \mathcal{A}, F)$ of functions using a limit process. Every function $f$ in $\mathcal{L}_\mu(X, \mathcal{A}, F)$ is the limit pointwise a.e. of a $N_1$-Cauchy sequence $(f_n)_{n \geq 1}$ (called an approximation sequence) of functions $f_n$ in $\text{Step}_\mu(X, \mathcal{A}, F)$. We also define the space $\mathcal{M}_\mu(X, \mathcal{A}, F)$ of $\mu$-measurable functions, and $\mathcal{L}_\mu(X, \mathcal{A}, F)$ is the subspace of $\mathcal{M}_\mu(X, \mathcal{A}, F)$ consisting of the functions for which the integral is well defined.

It turns out that $\mathcal{L}_\mu(X, \mathcal{A}, F)$ is complete with respect to an extension $\| \|_1$ of the semi-norm $N_1$, and the integral $\int f \, d\mu$ of any function $f \in \mathcal{L}_\mu(X, \mathcal{A}, F)$ can be defined by a limit process. There are technical complications when $F$ is infinite-dimensional, and it also takes some work to show that the integral of a function $f \in \mathcal{L}_\mu(X, \mathcal{A}, F)$ does not depend on the approximation sequence used to define $f$, but all difficulties can be overcome. Finally, the subspace $\mathcal{N}$ of functions $f$ such that $\|f\|_1 = 0$ is the set of functions equal to 0 a.e., and we obtain the complete space $(L_\mu(X, \mathcal{A}, F), \| \|_1)$ of
the first approach as the quotient space $\mathcal{L}_\mu(X, \mathcal{A}, F)/\mathcal{N}$. However, the construction of $\mathcal{L}_\mu(X, \mathcal{A}, F)$ is much more informative.

We also investigate convergence properties of $\mathcal{L}_\mu(X, \mathcal{A}, F)$, as well as other related spaces (the spaces $\mathcal{L}^p_\mu(X, \mathcal{A}, F)$, $p = 1, 2, \infty$). We conclude with the construction of the integral on a product space.

The vector valued-integral defined in this chapter (where the space $F$ of values is a Banach space) was first discovered by Bochner in 1933. The version discussed here is due to Dunford (1935), and is presented in detail in Dunford and Schwartz [16]. More recent expositions of this method are given in Lang [24] and Marle [27].

4.1 Measurable Maps

Measurable functions are functions between measurable spaces that are the analog of continuous functions between topological spaces, but as we will see, they are a lot more flexible, especially in terms of convergence properties. In this chapter, our presentation follows Marle [27] and Lang [24] very closely.

Definition 4.1. Given any two measurable spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$, a function $f: X \to Y$ is measurable if $f^{-1}(B) \in \mathcal{A}$ for every $B \in \mathcal{B}$. A measurable function is also called a measurable map.

If $(X, \mathcal{A})$ is a measurable space, then obviously the identity $\text{id}: X \to X$ is measurable.

The composition of two measurable maps is also measurable.

Proposition 4.1. Given three measurable spaces $(X, \mathcal{A})$, $(Y, \mathcal{B})$, and $(Z, \mathcal{C})$, if $f: X \to Y$ and $g: Y \to Z$ are measurable maps, then $g \circ f: X \to Z$ is a measurable map.

Proof. Recall that one of the properties of inverse images is that $(g \circ f)^{-1}(C) = f^{-1}(g^{-1}(C))$, for any subset $C$ of $Z$. But if $C \in \mathcal{C}$, since $g$ is measurable, $g^{-1}(C) \in \mathcal{B}$, and since $f$ is measurable, $f^{-1}(g^{-1}(C)) \in \mathcal{A}$, which shows that $g \circ f$ is measurable.

Remark: The above properties show that measurable spaces are the objects of a category whose morphisms are the measurable maps.

Proposition 4.2. Let $X$ and $Y$ be any two nonempty sets, and let $f: X \to Y$ be a function between them.

(1) If $\mathcal{A}$ is a $\sigma$-algebra on $X$, then we can define $\mathcal{A}_f$ as the family of subsets of $Y$ given by

$$\mathcal{A}_f = \{ B \in 2^Y | f^{-1}(B) \in \mathcal{A} \}.$$ 

Then $\mathcal{A}_f$ is the largest $\sigma$-algebra on $Y$ which makes $f$ measurable.
(2) If $\mathcal{B}$ is a $\sigma$-algebra on $Y$, then let $f^{-1}(\mathcal{B})$ be the family of subsets of $X$ given by

$$f^{-1}(\mathcal{B}) = \{f^{-1}(B) \in 2^X \mid B \in \mathcal{B}\}.$$  

Then $f^{-1}(\mathcal{B})$ is the smallest $\sigma$-algebra on $X$ which makes $f$ measurable.

The proof of Proposition 4.2 is left as an exercise.

Using Proposition 4.2 we obtain the following proposition which gives simple criteria to check that a map is measurable.

**Proposition 4.3.** Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be two measurable spaces.

1. If $S$ generates the $\sigma$-algebra $\mathcal{B}$ (which means that the smallest $\sigma$-algebra containing $S$ is $\mathcal{B}$), then a function $f : X \to Y$ is measurable iff $f^{-1}(S) \in \mathcal{A}$ for all $S \in S$.

2. If $Y$ is a topological space and if $\mathcal{B}$ is its Borel $\sigma$-algebra of open subsets, then a function $f : X \to Y$ is measurable iff $f^{-1}(U) \in \mathcal{A}$ for every open subset $U$ of $Y$ (or $f^{-1}(U) \in \mathcal{A}$ for every closed subset $U$ of $Y$).

3. If $X$ and $Y$ are both topological spaces and if $\mathcal{A}$ and $\mathcal{B}$ are their respective Borel $\sigma$-algebras, then every continuous map $f : X \to Y$ is measurable.

Given any subset $A$ of $X$, recall that the characteristic function $\chi_A$ of $A$ is defined by

$$\chi_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A. \end{cases}$$

Then, as illustrated in Figure 4.1, it is easy to show that for any subset $A$ of $X$, the function $\chi_A : X \to \mathbb{R}$ (where $\mathbb{R}$ is equipped with its $\sigma$-algebra of Borel sets) is measurable iff $A \in \mathcal{A}$, that is, $A$ is measurable.

In the theory of integration, all maps of interest will be measurable maps$^1$ $f : X \to F$ where $(X, \mathcal{A})$ is a measurable space, and $(F, \mathcal{B})$ is a measurable space such that either $F = \mathbb{R}$, or $F = \mathbb{C}$, or more generally $F$ is a Banach space (a complete normed vector space over $\mathbb{R}$ or $\mathbb{C}$), and $\mathcal{B}$ is the Borel $\sigma$-algebra of open subsets of $F$. In this case, various operations can be performed on functions $f : X \to F$.

Assume that $F$ is a normed vector space over the field $K$, where $K = \mathbb{R}$ or $K = \mathbb{C}$, and that $f : X \to F$ is any function, not necessarily measurable.

1. Given any function $f : X \to F$, for any $\lambda \in K$, let $\lambda f : X \to F$ be the function given by

$$(\lambda f)(x) = \lambda f(x), \quad x \in X.$$  

$^1$Actually, not quite in the most general case, but they will be equal to a measurable map a.e.
4.1. MEASURABLE MAPS

Figure 4.1: The upper figure illustrates $\chi_A: X \to \mathbb{R}$. If $S_1 \subset \mathbb{R}$ contains 1 but not 0, $\chi_A^{-1}(S_1) = A$. If $S_2 \subset \mathbb{R}$ contains 0 but not 1, $\chi_A^{-1}(S_1) = X - A$. Finally, if $S_3 \subset \mathbb{R}$ contains both 0 and 1, $\chi_A^{-1}(S_1) = A \cup (X - A) = X$.

2. Given any function $f: X \to F$, let $\|f\|: X \to \mathbb{R}^+$ be the function given by

$$\|f\|(x) = \|f(x)\|, \quad x \in X.$$ 

Beware that $\|f\|$ is not the norm of the function $f$, where $\|\|$ is the norm on some function space consisting of functions from $X$ to $F$. Instead, $\|f\|$ is the function defined pointwise as $\|f(x)\|$ for every $x \in X$, where $\|f(x)\|$ is the norm of $f(x)$ in $F$. This notation is somewhat confusing but appears to be standard. Later on, we will equip our space of functions from $X$ to $F$ with a norm, but it will be denoted $\|\|_1$, or more generally $\|\|_p$, so there will be no risk of confusion.

3. For any two functions $f: X \to \mathbb{R}$ and $g: X \to \mathbb{R}$, let $\sup(f, g)$ and $\inf(f, g)$ be the functions given by

$$\sup(f, g)(x) = \max(f(x), g(x)), \quad x \in X,$n
$$\inf(f, g)(x) = \min(f(x), g(x)), \quad x \in X.$$
4. For any two functions \( f: X \to \mathbb{R} \), let \( f^+ \) and \( f^- \) be the functions given by

\[
\begin{align*}
f^+(x) &= \begin{cases} 
0 & \text{if } f(x) \leq 0 \\
f(x) & \text{if } f(x) > 0,
\end{cases} \\
f^-(x) &= \begin{cases} 
0 & \text{if } f(x) \geq 0 \\
-f(x) & \text{if } f(x) < 0.
\end{cases}
\end{align*}
\]

We also define \( |f| = f^+ + f^- = \sup(f, -f) \). Observe that \( f = f^+ - f^- \).

5. For any two functions \( f: X \to F \) and \( g: X \to F \), let \( f + g: X \to F \) be the function given by

\[
(f + g)(x) = f(x) + g(x), \quad x \in X.
\]

6. For any two functions \( f: X \to K \) and \( g: X \to K \), where \( K = \mathbb{R} \) or \( K = \mathbb{C} \), let \( fg: X \to K \) be the function given by

\[
(fg)(x) = f(x)g(x), \quad x \in X.
\]

**Definition 4.2.** Let \((X, \mathcal{A})\) be a measurable space, and let \((F, \mathcal{B})\) be a measurable space such that either \( F = \mathbb{R} \), or \( F = \mathbb{C} \), or more generally \( F \) is a Banach space (a complete normed vector space over \( \mathbb{R} \) or \( \mathbb{C} \)), and \( \mathcal{B} \) is the Borel \( \sigma \)-algebra of open subsets of \( F \). The set of measurable maps \( f: X \to F \) is denoted by \( \mathcal{M}(X, \mathcal{A}, F) \).

The following technical result is needed.

**Proposition 4.4.** Let \((X, \mathcal{A})\) be any measurable space, and assume that \( F \) is a normed vector space over the field \( K \), where \( K = \mathbb{R} \) or \( K = \mathbb{C} \). The following properties hold:

1. For any \( f \in \mathcal{M}(X, \mathcal{A}, F) \) and any \( \lambda \in K \), we have \( \lambda f \in \mathcal{M}(X, \mathcal{A}, F) \).

2. For any \( f \in \mathcal{M}(X, \mathcal{A}, F) \), we have \( \|f\| \in \mathcal{M}(X, \mathcal{A}, \mathbb{R}) \).

3. For any \( f \in \mathcal{M}(X, \mathcal{A}, \mathbb{R}) \) and any \( g \in \mathcal{M}(X, \mathcal{A}, \mathbb{R}) \), we have \( \sup(f, g), \inf(f, g), f^+, f^- \), \( |f| \in \mathcal{M}(X, \mathcal{A}, \mathbb{R}) \).
4. For any \( f \in \mathcal{M}(X, A, F) \) and any \( g \in \mathcal{M}(X, A, F) \), if \( f(X) \) and \( g(X) \) are separable subsets of \( F \), then \( f + g \in \mathcal{M}(X, A, F) \). In particular, if \( F \) is separable, then \( \mathcal{M}(X, A, F) \) is a vector space over \( K \).

5. For any \( f \in \mathcal{M}(X, A, K) \) and any \( g \in \mathcal{M}(X, A, K) \), we have \( fg \in \mathcal{M}(X, A, K) \). This implies that \( \mathcal{M}(X, A, K) \) is actually a \( K \)-algebra.

One will observe that in (4), if \( F \) is infinite-dimensional, the sum of two measurable maps may not be measurable. This is the first technical difficulty of the general theory of integration (with values in an infinite-dimensional vector space). As we will see, a second technical difficulty has to do with the approximation of a measurable map by step functions. Fortunately these technical difficulties can be overcome in a simple way.

The following important result shows that measurable maps behave better than continuous maps in terms of simple convergence.

**Theorem 4.6.** Let \( (X, A) \) and \( (F, B) \) be two measurable spaces, where \( F \) is a metric space and \( B \) is the Borel \( \sigma \)-algebra on \( F \). If \( (f_n)_{n \geq 1} \) is a sequence of measurable maps \( f_n \in \mathcal{M}(X, A, F) \) which converges pointwise to a function \( f : X \to F \), then \( f \in \mathcal{M}(X, A, F) \); that is, \( f \) is measurable.

A proof of Theorem 4.6 can be found in Lang [24] (Chapter VI, Section 1, Property M7).

Our next goal is to generalize the notion of step function given in Definition 1.10 to the framework of measure spaces.

### 4.2 Step Maps on a Measurable Space

Let \( (X, A) \) be a measurable space. The generalization of the notion of step map is obtained by replacing the intervals \((a_i, a_{i+1})\) by arbitrary measurable sets.

**Definition 4.3.** Let \( (X, A) \) be a measurable space, and let \( F \) be any set. A function \( f : X \to F \) is a step map (with respect to \( A \)) if there is a finite partition \((A_1, \ldots, A_n)\) of \( X \) by pairwise disjoint nonempty subsets \( A_i \in A \) such that \( X = \bigcup_{i=1}^n A_i \), and such that the restriction of \( f \) to each \( A_i \) is a constant function with some value \( y_i \in F \). The partition \((A_1, \ldots, A_n)\) is said to be adapted to \( f \). See Figure 4.2. The partition \((A_1, \ldots, A_n)\) is said to be adapted to \( f \). The set of all step maps is denoted by \( \text{Step}(X, A, F) \).

Observe that every constant function is a step function, and that \( f(X) \) is a finite subset of \( F \). At this stage, no measure \( \mu \) is involved, but for the theory of integration, we will have a measure space \((X, A, \mu)\) and we will need to require each \( A_i \) for which \( y_i \neq 0 \) to have finite measure (this makes sense since in this case \( F \) is a vector space).

We gather some useful properties of step functions in the following proposition.
Figure 4.2: Let \((X, \mathcal{A}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))\) and \(F = \mathbb{R}\). A step map is shown in blue with values \(\{y_i\}_{i=1}^n\). The partition \((A_1, \ldots, A_n)\) adapted to \(f\) is shown underneath the peach box.

**Proposition 4.7.** Let \((X, \mathcal{A})\) be a measurable space, and let \(F\) be any set.

1. For any \(\sigma\)-algebra \(\mathcal{B}\) on \(F\), every step map \(\text{Step}(X, \mathcal{A}, F)\) is measurable.

2. Let \(F_1, F_2, G\) be three sets, and let \(h: F_1 \times F_2 \to G\) be any function. For any \(f_1 \in \text{Step}(X, \mathcal{A}, F_1)\) and any \(f_2 \in \text{Step}(X, \mathcal{A}, F_2)\), we have \(h \circ (f_1, f_2) \in \text{Step}(X, \mathcal{A}, G)\).

3. If \(K = \mathbb{R}\) or \(K = \mathbb{C}\), then \(\text{Step}(X, \mathcal{A}, K)\) is a vector space and a ring under pointwise multiplication of functions. Thus, \(\text{Step}(X, \mathcal{A}, K)\) is an algebra over \(K\).

4. If \(F\) is a vector space over \(K\) (with \(K = \mathbb{R}\) or \(K = \mathbb{C}\)), then \(\text{Step}(X, \mathcal{A}, F)\) is a vector space over \(K\), and a module over \(\text{Step}(X, \mathcal{A}, K)\), which means that if \(f \in \text{Step}(X, \mathcal{A}, F)\) and \(g \in \text{Step}(X, \mathcal{A}, K)\), then \(gf \in \text{Step}(X, \mathcal{A}, F)\).

5. If \(F\) is a normed vector space, and if \(f \in \text{Step}(X, \mathcal{A}, F)\), then \(\|f\| \in \text{Step}(X, \mathcal{A}, \mathbb{R})\).

6. If \(f \in \text{Step}(X, \mathcal{A}, \mathbb{R})\) and \(g \in \text{Step}(X, \mathcal{A}, \mathbb{R})\), then we have \(\sup(f, g), \inf(f, g), f^+, f^-, |f| \in \text{Step}(X, \mathcal{A}, \mathbb{R})\).
By Theorem 4.6, if $F$ is a metric space equipped with its $\sigma$-algebra of Borel sets, every sequence $(f_n)_{n \geq 1}$ of step functions $f_n \in \text{Step}(X, A, F)$ that converges pointwise to a function $f : X \to F$ must be measurable, since the $f_n$ are measurable by Proposition 4.7.

Unfortunately, in general, a measurable map $f : X \to F$ may not be the pointwise limit of a sequence of step maps if $F$ has infinite dimension. For one thing, such a limit of steps maps has its image contained in the closure of a countable subset of $F$. This is the second technical difficulty of the general theory.

To overcome this second difficulty, one needs to define a more refined notion of measurable map and of step map. We will do so shortly, but first we observe that if we only need to consider values in a finite-dimensional vector space, then there is no problem.

**Proposition 4.8.** Let $(X, A)$ and $(F, B)$ be two measurable spaces, where $F$ is a topological space and $B$ is its Borel $\sigma$-algebra, and let $f : X \to F$ be a measurable map.

1. If $F$ is either a finite-dimensional vector space over $\mathbb{R}$ or $\mathbb{C}$, or $F = \mathbb{R}_+$, then there is a sequence $(f_n)$ of step maps $f_n \in \text{Step}(X, A, F)$ that converges pointwise to $f$. If $F = \mathbb{R}_+$, we may assume that the $f_n$ take finite values.

2. If $F = \mathbb{R}$ or $F = \mathbb{R}_+$, and if $f \geq 0$, then we may assume that $f_n \geq 0$ and $f_n \leq f_{n+1}$ for all $n \geq 1$.

A proof of Proposition 4.8 can be found in Lang [24] (Chapter VI, Section 1, Properties M8 and M9).

### 4.3 $\mu$-Measurable Maps and $\mu$-Step Maps

We explained in the previous section that in general, the space $\mathcal{M}(X, A, F)$ of measurable maps from $X$ to $F$ is not a vector space, and that a measurable map $f : X \to F$ may not be the pointwise limit of a sequence of step maps. This suggests modifying the notion of measurable map and the notion of step map to recover these properties. The second property is crucial in extending the notion of integral to more general functions.

So far, the space $X$ was only a measurable space, but no measure was involved. The new ingredient is to define a suitable notions of step map and measurable map relative to a measure space $(X, A, \mu)$, where the measure $\mu$ plays a role.

The main trick is to relax the notion of pointwise convergence to pointwise convergence almost everywhere, and more generally, to consider that two functions are equivalent if they are equal almost everywhere (they differ on a null set). The plan is the following:

1. Define the space $\text{Step}_\mu(X, A, F)$ of $\mu$-step maps.

2. Define the space $\mathcal{M}_\mu(X, A, F)$ of $\mu$-measurable maps, where a $\mu$-measurable map is the limit of a sequence $(f_n)$ of $\mu$-step maps $f_n \in \text{Step}_\mu(X, A, F)$ converging pointwise almost everywhere.
3. Prove that if $F$ is a vector space, then $\mathcal{M}_\mu(X, \mathcal{A}, F)$ is a vector space.

Our presentation of the method that we just sketched follows Marle [27] and Lang [24] very closely. It is a generalization (with some simplifications) to functions with values in a Banach space of the approach followed by Halmos [22]. The results that we state without proof are proved either in Marle [27] or in Lang [24].

**Definition 4.4.** Let $(X, \mathcal{A}, \mu)$ be a measure space, and let $F$ be any vector space (over $\mathbb{R}$ or $\mathbb{C}$). A function $f: X \to F$ is a $\mu$-step map if it is a step map, and if $\{x \in X \mid f(x) \neq 0\} \in \mathcal{A}$ and has finite measure. See Figure 4.3. The set of $\mu$-step maps is denoted by $\text{Step}_\mu(X, \mathcal{A}, F)$.

![Figure 4.3](image)

Figure 4.3: Let $(X, \mathcal{A}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $F = \mathbb{R}$. A $\mu$-step map is shown in red where $A_1 \cup A_2 \cup A_3 = \{x \in X \mid f(x) \neq 0\} \in \mathcal{A}$.

For technical reasons, it is useful to have the following equivalent characterization of a $\mu$-step map.

**Proposition 4.9.** Let $(X, \mathcal{A}, \mu)$ be a measure space, and let $F$ be any vector space (over $\mathbb{R}$ or $\mathbb{C}$). A function $f: X \to F$ is a $\mu$-step map iff there is a nonempty subset $A \in \mathcal{A}$ of finite measure such that $f$ vanishes outside $A$, that is, $f(x) = 0$ for all $x \in X - A$, and if there is a finite partition $(A_1, \ldots, A_n)$ of $A$ of subsets $A_i \in \mathcal{A}$ (nonempty pairwise disjoint subsets) such that the restriction of $f$ to each $A_i$ has a constant value $y_i$.

**Proof.** Let $f$ be a step map with respect to a partition $(A_1, \ldots, A_n)$ of $X$ such that $\{x \in X \mid f(x) \neq 0\} \in \mathcal{A}$ and has finite measure. Then any $A_i$ on which $f$ has value $y_i \neq 0$ must have finite measure. If $f = 0$ on $X$, then pick $A$ to be any $A_i$ and the partition to be $(A_i)$. Otherwise, let $J = \{j \in \{1, \ldots, n\} \mid f \neq 0 \text{ on } A_j\}$, and let $A = \bigcup_{j \in J} A_j$. Then, $(A_j)_{j \in J}$ is a partition of $A$ with $A_j \in \mathcal{A}$, where $A$ is a nonempty set of finite measure, and $f$ vanishes on $X - A$. See Figure 4.4.
4.3. $\mu$-MEASURABLE MAPS AND $\mu$-STEP MAPS

Figure 4.4: Let $(X, \mathcal{A}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $F = \mathbb{R}$. A step map is shown in red with adapted partition $\mathbb{R} = \bigcup_{i=1}^7 A_i$. To interpret this step map as a $\mu$-step map, let $A = A_2 \cup A_4 \cup A_6$, where $\mu(A) < \infty$, $A = \{x \in X \mid f(x) \neq 0\} \in \mathcal{A}$. Then $f(x) = 0$ on $X - A$ where $X - A = A_1 \cup A_3 \cup A_5 \cup A_7$.

Conversely, since $A$ has finite measure and since the $A_i$ belongs to $\mathcal{A}$, each $A_i$ has finite measure, so $\{x \in X \mid f(x) \neq 0\} \in \mathcal{A}$ is a set of finite measure. If $A = X$, then we already have a step map (as defined in Definition 4.3). Otherwise, $X - A \in \mathcal{A}$ and $f$ vanishes on $X - A$, so $(A_1, \ldots, A_n, X - A)$ is partition of $X$, and $f$ is a step map with respect to this partition. See Figure 4.5.

The condition that a $\mu$-step map must vanish outside of a measurable set of finite measure is the measure-theoretic analog of the topological notion of compact support.

Proposition 4.9 suggests the following equivalent definition of a $\mu$-step map.

**Definition 4.5.** Let $(X, \mathcal{A}, \mu)$ be a measure space, and let $F$ be any vector space (over $\mathbb{R}$ or $\mathbb{C}$). A function $f: X \to F$ is a $\mu$-step map if there is a nonempty subset $A \in \mathcal{A}$ of finite measure such that $f$ vanishes outside $A$, that is, $f(x) = 0$ for all $x \in X - A$, and if there is a finite partition $(A_1, \ldots, A_n)$ of $A$ consisting of nonempty pairwise disjoint subsets in $\mathcal{A}$, such that the restriction of $f$ to each $A_i$ has a constant value $y_i$ (possibly zero). The partition $(A_1, \ldots, A_n)$ of $A$ is said to be adapted to $f$.

Technically, Definition 4.5 appears to be more convenient. Observe that a $\mu$-step map can be expressed as a (necessarily finite) linear combination

$$f = \sum_{i=1}^n y_i \chi_{A_i},$$

for some $y_i \in F$ and for some nonempty pairwise disjoint measurable sets $A_i \in \mathcal{A}$ of finite measure, a concise and convenient representation.
CHAPTER 4. INTEGRATION

Figure 4.5: Let \((X, \mathcal{A}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))\) and \(F = \mathbb{R}\). A \(\mu\)-step map is shown in red with \(A = A_1 \cup A_2 \cup A_3 \cup A_4\). The turquoise set is \(X - A\) and \((A_1, A_2, A_3, A_4, X - A)\) forms an adapted partition for the corresponding step map.

**Remark:** The proof of Proposition 4.9 shows that if a \(\mu\)-step function \(f\) is not identically zero, then we can find a subset \(A\) in \(\mathcal{A}\) of finite measure, and a partition \((A_1, \ldots, A_n)\) of \(A\) of subsets in \(\mathcal{A}\) such that the value of \(f\) on each \(A_i\) is nonzero, and \(f\) is zero outside of \(A\). However, it turns out to be more convenient for certain proofs to allow \(f\) to be zero on some of the \(A_i\), and this is why we allow this possibility in Definition 4.5.

**Example 4.1.** Consider the function \(f: \mathbb{R} \to \mathbb{R}\), given by

\[
\begin{aligned}
f(x) &= \begin{cases} 
0 & \text{if } x < 0 \text{ or } x > 1 \\
1 & \text{if } x \in [0, 1/2) - \mathbb{Q} \\
0 & \text{if } x \in [0, 1/2] \cap \mathbb{Q} \\
2 & \text{if } x \in [1/2, 1) - \mathbb{Q} \\
0 & \text{if } x \in [1/2, 1] \cap \mathbb{Q}.
\end{cases}
\end{aligned}
\]

See Figure 4.6 If we let \(A_1 = [0, 1/2) - \mathbb{Q}\), \(A_2 = [0, 1/2] \cap \mathbb{Q}\), \(A_3 = [1/2, 1) - \mathbb{Q}\), \(A_4 = [1/2, 1] \cap \mathbb{Q}\), and \(A = [0, 1]\), with the Lebesgue measure \(\mu_L\) on \(\mathbb{R}\), then \(A_1, A_2, A_3, A_4\) are Lebesgue measurable, \(\mu(A_1) = 1/2\), \(\mu(A_2) = 0\), \(\mu(A_3) = 1/2\), \(\mu(A_3) = 0\), \((A_1, A_2, A_3, A_4)\) is a partition of \(A\), a set of measure 1. Thus \(f\) is a \(\mu_L\)-step function.

This example shows that a \(\mu\)-step function can be very complicated, unlike the step functions of Definition 1.10.

**Proposition 4.10.** Let \((X, \mathcal{A}, \mu)\) be a measure space, and let \(F\) be any vector space

1. Let \(F_1, F_2, G\) be three Banach spaces over \(\mathbb{R}\) or \(\mathbb{C}\), and let \(h: F_1 \times F_2 \to G\) be any function. If \(h\) satisfies \(h(0, 0) = 0\), then for any \(f_1 \in \text{Step}_\mu(X, \mathcal{A}, F_1)\) and any \(f_2 \in \text{Step}_\mu(X, \mathcal{A}, F_2)\), we have \(h \circ (f_1, f_2) \in \text{Step}_\mu(X, \mathcal{A}, G)\).
2. If $K = \mathbb{R}$ or $K = \mathbb{C}$, then $\text{Step}_\mu(X, \mathcal{A}, K)$ is a subspace of $\text{Step}(X, \mathcal{A}, K)$, and for any $g \in \text{Step}(X, \mathcal{A}, K)$ and any $f \in \text{Step}_\mu(X, \mathcal{A}, K)$ we have $gf \in \text{Step}_\mu(X, \mathcal{A}, K)$. Thus $\text{Step}_\mu(X, \mathcal{A}, K)$ is an ideal in $\text{Step}(X, \mathcal{A}, K)$.

3. If $F$ is a vector space over $K$ (with $K = \mathbb{R}$ or $K = \mathbb{C}$), then $\text{Step}_\mu(X, \mathcal{A}, F)$ is a subspace of $\text{Step}(X, \mathcal{A}, F)$ and a module over $\text{Step}(X, \mathcal{A}, K)$, which means that if $f \in \text{Step}_\mu(X, \mathcal{A}, F)$ and $g \in \text{Step}(X, \mathcal{A}, K)$, then $gf \in \text{Step}_\mu(X, \mathcal{A}, F)$.

4. If $F$ is a normed vector space, and if $f \in \text{Step}_\mu(X, \mathcal{A}, F)$, then $\|f\| \in \text{Step}_\mu(X, \mathcal{A}, \mathbb{R})$. In fact, if $f = \sum_{i=1}^{n} y_i \chi_{A_i}$, then $\|f\| = \sum_{i=1}^{n} |y_i| \chi_{A_i}$.

5. If $f \in \text{Step}_\mu(X, \mathcal{A}, \mathbb{R})$ and $g \in \text{Step}_\mu(X, \mathcal{A}, \mathbb{R})$, then $\sup(f, g), \inf(f, g), f^+, f^-, |f| \in \text{Step}_\mu(X, \mathcal{A}, \mathbb{R})$.

We now come to the crucial notion of $\mu$-measurable map.

**Definition 4.6.** Let $(X, \mathcal{A}, \mu)$ be a measure space, and let $F$ be any vector space (over $\mathbb{R}$ or $\mathbb{C}$). A function $f: X \to F$ is a $\mu$-measurable if there is a sequence $(f_n)_{n \geq 1}$ of $\mu$-step maps $f_n \in \text{Step}_\mu(X, \mathcal{A}, F)$ which converges pointwise to $f$ almost everywhere. See Figures 4.7 and 4.8. Recall that this means that there is a null set $Z \subseteq X$ such that for every $x \in X - Z$, the sequence $(f_n(x))$ converges to $f(x)$. The set of $\mu$-measurable maps is denoted by $\mathcal{M}_\mu(X, \mathcal{A}, F)$.

Observe that a $\mu$-measurable map is not necessarily measurable, so $\mathcal{M}_\mu(X, \mathcal{A}, F)$ is not a subspace of $\mathcal{M}(X, \mathcal{A}, F)$. However, we will see shortly that a $\mu$-measurable map is equal to a measurable map almost everywhere, and this is good enough to construct the Lebesgue integral. The following proposition can be proved using Proposition 4.10 by passing to the limit (carefully).
Figure 4.7: Let $X = F = \mathbb{R}$. Assume $\mu$ is the Lebesgue measure on $\mathbb{R}$. The graph of the $\mu$-measurable $f$ is shown in the upper left corner. The middle figure illustrates the sequence $(f_n)_{n \geq 1}$ of $\mu$-step maps $f_n \in \text{Step}_\mu(X, A, F)$ which converges pointwise to $f$ almost everywhere.

**Proposition 4.11.** Let $(X, A, \mu)$ be a measure space, and let $F$ be any vector space

1. Let $F_1, F_2, G$ be three Banach spaces over $\mathbb{R}$ or $\mathbb{C}$, and let $h: F_1 \times F_2 \to G$ be any function. If $h$ satisfies $h(0,0) = 0$, then for any $f_1 \in \mathcal{M}_{\mu}(X, A, F_1)$ and any $f_2 \in \mathcal{M}_{\mu}(X, A, F_2)$, we have $h \circ (f_1, f_2) \in \mathcal{M}_{\mu}(X, A, G)$.

2. If $K = \mathbb{R}$ or $K = \mathbb{C}$, then $\mathcal{M}_{\mu}(X, A, K)$ is a vector space, and for all $f, g \in \mathcal{M}_{\mu}(X, A, K)$ we have $fg \in \mathcal{M}_{\mu}(X, A, K)$. Thus $\mathcal{M}_{\mu}(X, A, K)$ is an algebra over $K$. For any $g \in \mathcal{M}(X, A, K)$ and any $f \in \mathcal{M}_{\mu}(X, A, K)$ we have $gf \in \mathcal{M}_{\mu}(X, A, K)$.

3. If $F$ is a vector space over $K$ (with $K = \mathbb{R}$ or $K = \mathbb{C}$), then $\mathcal{M}_{\mu}(X, A, F)$ is a vector space over $K$ and a module over $\mathcal{M}(X, A, K)$, which means that if $f \in \mathcal{M}_{\mu}(X, A, F)$ and $g \in \mathcal{M}(X, A, K)$, then $gf \in \mathcal{M}_{\mu}(X, A, F)$. The space $\mathcal{M}_{\mu}(X, A, F)$ is also a module over $\mathcal{M}_{\mu}(X, A, K)$.

4. If $F$ is a normed vector space, and if $f \in \mathcal{M}_{\mu}(X, A, F)$, then $\|f\| \in \mathcal{M}_{\mu}(X, A, \mathbb{R})$.

5. If $f \in \mathcal{M}_{\mu}(X, A, \mathbb{R})$ and $g \in \mathcal{M}_{\mu}(X, A, \mathbb{R})$, then we have $\text{sup}(f,g), \text{inf}(f,g), f^+, f^-, |f| \in \mathcal{M}_{\mu}(X, A, \mathbb{R})$.

The following result gives a characterization of a $\mu$-measurable map which shows that
a $\mu$-measurable map is equal to a measurable map almost everywhere, and that there are strong countability restrictions on its domain and their range.

**Proposition 4.12.** Let $(X, \mathcal{A}, \mu)$ be a measure space, and let $F$ be any Banach space. A function $f : X \to F$ is $\mu$-measurable iff there is a null set $Z$ such that the following three conditions hold:

1. There is a measurable map $g \in \mathcal{M}(X, \mathcal{A}, F)$ such that $f$ and $g$ are equal on $X - Z$.
2. The function $f$ vanishes outside of a measurable $\sigma$-finite subset of $X$.
3. The image $f(X - Z)$ is separable in $F$, which means that $f(X - Z)$ contains a countable dense subset.

In particular, if $\mu$ is $\sigma$-finite and if $F$ is separable, then $f : X \to F$ is $\mu$-measurable iff $f$ is measurable almost everywhere (there is a null set $Z$ such that $f$ agrees with a measurable map on $X - Z$).

A proof of Proposition 4.12 can be found in Lang [24] (Chapter VI, Section 1, Property M11). Again, Condition (2) is a measure-theoretic analog of the notion of compact support.

The version of Theorem 4.6 for $\mu$-measurable maps is stated below.

---

Figure 4.8: A more detailed look at the pointwise convergence of converges the $(f_n)_{n \geq 1}$. The pointwise limit only differs from $f$ on a set of measure zero, namely $\{3/2\}$. 
Theorem 4.13. Let \((X, \mathcal{A}, \mu)\) be a measure space and let \((F, \mathcal{B})\) be a measurable space, where \(F\) is a metric space and \(\mathcal{B}\) is the Borel \(\sigma\)-algebra on \(F\). If \((f_n)_{n \geq 1}\) is a sequence of \(\mu\)-measurable maps \(f_n \in M_\mu(X, \mathcal{A}, F)\) which converges pointwise to a function \(f : X \to F\), then \(f \in M_\mu(X, \mathcal{A}, F)\); that is, \(f\) is \(\mu\)-measurable.

A proof of Theorem 4.6 can be found in Lang [24] (Chapter VI, Section 1, Property M12).

We are now ready to construct a very general version of the integral. The original construction was first proposed by Lebesgue, but the more general version presented here applying to functions with values in a Banach space is due to Bochner and Dunford.

4.4 The Integral of \(\mu\)-Step Maps

Let \((X, \mathcal{A}, \mu)\) be a measure space and let \((F, \mathcal{B})\) be a measurable space consisting of a Banach space \(F\) and its Borel \(\sigma\)-algebra \(\mathcal{B}\). There is an “obvious” definition of the integral of a \(\mu\)-step map \(f = \sum_{i=1}^{n} y_i \chi_{A_i}\) (where \(y_i \in F\)), namely

\[
I(f) = \int f d\mu = \sum_{i=1}^{n} \mu(A_i)y_i.
\]

Since by definition the \(A_i\) belong to \(\mathcal{A}\) and have finite measure, the linear combination \(\sum_{i=1}^{n} \mu(A_i)y_i\) is a well-defined vector in \(F\). The only problem is that \(I(f)\) seems to depend on the subset \(A\) (and its partition) chosen to express \(f\), but it is easy to show that \(I(f)\) is independent of the representation of \(f\). Then it is easy to show that \(I : \text{Step}_\mu(X, \mathcal{A}, F) \to F\) is a linear map. Furthermore, by Proposition 4.11, we have \(\|f\| \in \text{Step}_\mu(X, \mathcal{A}, \mathbb{R})\), so we can define

\[
N_1(f) = \int \|f\| d\mu,
\]

and we have

\[
\left\| \int f d\mu \right\| \leq \int \|f\| d\mu = N_1(f).
\]

It turns out that \(N_1\) satisfies all the axioms of a norm, except that \(N_1(f) = 0\) does not necessarily imply that \(f = 0\). We say that \(N_1\) is a semi-norm, see Definition 13.3. Fortunately, for any \(f \in \text{Step}_\mu(X, \mathcal{A}, F)\), we have \(N_1(f) = 0\) iff \(f = 0\), except on a subset of measure zero.

We can define the notion of \(N_1\)-Cauchy sequence of a sequence \((f_n)\) of functions \(f_n \in \text{Step}_\mu(X, \mathcal{A}, F)\) as follows: for all \(\epsilon > 0\), there is some \(N > 0\), such that for all \(m, n \geq N\), we have \(N_1(f_m - f_n) < \epsilon\). We can also define the notion of \(N_1\)-convergence of a sequence \((f_n)\) of functions \(f_n \in \text{Step}_\mu(X, \mathcal{A}, F)\) to a limit \(f \in \text{Step}_\mu(X, \mathcal{A}, F)\) as follows: for all \(\epsilon > 0\), there is some \(N > 0\), such that for all \(n \geq N\), we have \(N_1(f - f_n) < \epsilon\). A convergent \(N_1\)-sequence does not necessarily have a unique limit, but we will see that any two limits are equal a.e.
4.4. THE INTEGRAL OF \( \mu \)-STEP MAPS

The problem is that an \( N_1 \)-Cauchy sequence may not have a limit in \( \text{Step}_\mu(X, \mathcal{A}, F) \). Thus we are led to completing \( \text{Step}_\mu(X, \mathcal{A}, F) \) with respect to the semi-norm \( N_1 \). This can be done and we obtain a vector space \( \mathcal{L}_\mu(X, \mathcal{A}, F) \) which is a subspace of \( \mathcal{M}_\mu(X, \mathcal{A}, F) \). The integral map \( I \) and the semi-norm \( N_1 \) can be extended to \( \mathcal{L}_\mu(X, \mathcal{A}, F) \) as a semi-norm denoted \( \| \cdot \|_1 \), the space \( \mathcal{L}_\mu(X, \mathcal{A}, F) \) is Cauchy-complete with respect to the semi-norm \( \| \cdot \|_1 \), and \( \text{Step}_\mu(X, \mathcal{A}, F) \) is dense in \( \mathcal{L}_\mu(X, \mathcal{A}, F) \) with respect to the semi-norm \( \| \cdot \|_1 \).

It also turns out that the subspace \( SN \) of \( \text{Step}_\mu(X, \mathcal{A}, F) \) consisting of all functions \( f \) such that \( N_1(f) = 0 \) is the set of functions in \( \text{Step}_\mu(X, \mathcal{A}, F) \) that are equal to 0 a.e. Similarly, the subspace \( \mathcal{N} \) of \( \mathcal{L}_\mu(X, \mathcal{A}, F) \) consisting of all functions \( f \) such that \( \| f \|_1 = 0 \) is the set of functions in \( \mathcal{L}_\mu(X, \mathcal{A}, F) \) that are equal to 0 a.e. Thus, we can form the quotients spaces \( \text{Step}_\mu(X, \mathcal{A}, F) = \text{Step}_\mu(X, \mathcal{A}, F)/SN \) and \( \mathcal{L}_\mu(X, \mathcal{A}, F) = \mathcal{L}_\mu(X, \mathcal{A}, F)/\mathcal{N} \).

In \( \text{Step}_\mu(X, \mathcal{A}, F) \) and in \( \mathcal{L}_\mu(X, \mathcal{A}, F) \) the semi-norm \( \| \cdot \|_1 \) is really a norm, and \( \mathcal{L}_\mu(X, \mathcal{A}, F) \) is the completion of \( \text{Step}_\mu(X, \mathcal{A}, F) \).

Theoretically, we could define \( \mathcal{L}_\mu(X, \mathcal{A}, F) \) directly as the Cauchy completion (see Theorem 13.58 and Theorem 13.68) of \( \text{Step}_\mu(X, \mathcal{A}, F) \), but we obtain equivalence classes of Cauchy sequences of equivalence classes of functions in \( \text{Step}_\mu(X, \mathcal{A}, F) \), which are not easily interpretable as functions. The same space \( \mathcal{L}_\mu(X, \mathcal{A}, F) \) is obtained, see the diagram below.

\[
\text{Step}_\mu(X, \mathcal{A}, F) \xrightarrow{\text{completion}} \mathcal{L}_\mu(X, \mathcal{A}, F) \\
\text{Step}_\mu(X, \mathcal{A}, F) = \text{Step}_\mu(X, \mathcal{A}, F)/SN \xrightarrow{\text{completion}} \mathcal{L}_\mu(X, \mathcal{A}, F) = \mathcal{L}_\mu(X, \mathcal{A}, F)/\mathcal{N}.
\]

The construction that we alluded to, although involving some extra work, yields a very clear description of these equivalence classes in terms of functions (in \( \mathcal{L}_\mu(X, \mathcal{A}, F) \)). The completeness of \( \mathcal{L}_\mu(X, \mathcal{A}, F) \) (under the \( \| \cdot \|_1 \)-norm) is also immediately obtained.

As in the previous section the results that we state without proof are proved either in Marle [27] or in Lang [24].

We now return to the definition of the integral of a \( \mu \)-step maps.

**Proposition 4.14.** Let \( (X, \mathcal{A}, \mu) \) be a measure space and let \((F, \mathcal{B})\) be a measurable space, with \( F \) a Banach space and \( \mathcal{B} \) its Borel \( \sigma \)-algebra. For any \( \mu \)-step map \( f \in \text{Step}_\mu(X, \mathcal{A}, F) \), for any two partitions \((A_1, \ldots, A_m)\) and \((B_1, \ldots, B_n)\) adapted to \( f \), so that \( f = \sum_{i=1}^{m} y_i \chi_{A_i} = \sum_{j=1}^{n} z_j \chi_{B_j} \), we have

\[
\sum_{i=1}^{m} \mu(A_i) y_i = \sum_{j=1}^{n} \mu(B_j) z_j.
\]

Proposition 4.14 justifies the following definition.
Definition 4.7. Let \((X, \mathcal{A}, \mu)\) be a measure space and let \((F, \mathcal{B})\) be a measurable space, with \(F\) a Banach space and \(\mathcal{B}\) its Borel \(\sigma\)-algebra. For any \(\mu\)-step map \(f \in \text{Step}_\mu(X, \mathcal{A}, F)\), the common value
\[
\int f \, d\mu
\]
of the expression
\[
I(f) = \sum_{i=1}^{n} \mu(A_i) y_i
\]
for any partition \((A_1, \ldots, A_n)\) adapted to \(f\) is called the integral of \(f\) (relative to the measure \(\mu\)).

Recall that if the \(\mu\)-step map \(f\) is expressed as \(f = \sum_{i=1}^{n} y_i \chi_{A_i}\), then the \(\mu\)-set map \(\|f\| = \sum_{i=1}^{n} \|y_i\| \chi_{A_i}\). We also define the semi-norm \(N_1(f)\) of \(f\) as
\[
N_1(f) = \int \|f\| \, d\mu = \sum_{i=1}^{n} \mu(A_i) \|y_i\|.
\]

For any measurable subset \(E \in \mathcal{A}\), since \(\chi_E f \in \text{Step}_\mu(X, \mathcal{A}, F)\), we let
\[
\int_E f \, d\mu = \int \chi_E f \, d\mu.
\]

For simplicity of notation, we often write \(\int_E f\) instead of \(\int_E f \, d\mu\), and if \(E = X\), we write \(\int f\) instead of \(\int f \, d\mu\).

We stress that the integral \(\int f \, d\mu\) or \(\int_E f \, d\mu\) is always finite; that is, an element of \(F\), but not \(\infty\). This is in contrast with the approach where the integral of a step function may have the value \(+\infty\), as in Rudin [36] (Chapter 1). At some later stage, in defining the space \(L^1(X, \mathcal{A}, F)\), it is necessary to require the integral to be finite anyway. We find the approach where the integral is finite in the first place less confusing. It also yields a more explicit definition of \(L^1(X, \mathcal{A}, F)\).

Here are some of the main properties of the integral.

Proposition 4.15. Let \((X, \mathcal{A}, \mu)\) be a measure space and let \((F, \mathcal{B})\) be a measurable space, with \(F\) a Banach space and \(\mathcal{B}\) its Borel \(\sigma\)-algebra. The following properties hold:

1. The integral map \(\int : \text{Step}_\mu(X, \mathcal{A}, F) \to F\) is a linear map.

2. If \(A\) and \(B\) are any two disjoint measurable subsets, then
\[
\int_{A \cup B} f \, d\mu = \int_A f \, d\mu + \int_B f \, d\mu
\]

---

\(^2\)This integral is usually called the Lebesgue integral or Bochner integral. A more appropriate name might be the Bochner–Dunford integral.
4.4. THE INTEGRAL OF $\mu$-STEP MAPS

3. For any map $f \in \text{Step}_\mu(X, A, F)$, we have $\|f\| \in \text{Step}_\mu(X, A, \mathbb{R})$, and

$$\left| \int fd\mu \right| \leq \int \|f\| d\mu = N_1(f).$$

We also have

$$\int \|f\| d\mu \leq \mu(\{x \in X \mid f(x) \neq 0\}) \|f\|_\infty.$$

4. For any two maps $f, g \in \text{Step}_\mu(X, A, F)$, if $f = g$ a.e., then $\int fd\mu = \int gd\mu$.

5. For any two maps $f, g \in \text{Step}_\mu(X, A, F)$, if $f \leq g$ a.e., then $\int fd\mu \leq \int gd\mu$. In particular, if $f \geq 0$ a.e., then $\int fd\mu \geq 0$.

6. $N_1$ is a semi-norm on $\text{Step}_\mu(X, A, F)$. Furthermore, for any $f \in \text{Step}_\mu(X, A, F)$, we have $N_1(f) = 0$ iff $f = 0$, except on a subset of measure zero.

7. If $F_1$ and $F_2$ are two Banach spaces over $\mathbb{R}$ or $\mathbb{C}$, and if $f : F_1 \to F_2$ is a continuous linear map, then for any $f \in \text{Step}_\mu(X, A, F_1)$, we have $h \circ f \in \text{Step}_\mu(X, A, F_2)$, and

$$\int (h \circ f) d\mu = h \left( \int fd\mu \right)$$

If $F = \mathbb{C}$, the above property holds for any semi-linear map.

Proof. We prove (3) and (6), leaving the other properties as exercises.

(3) If $f = \sum_{i=1}^{n} y_i \chi_{A_i}$ then $\int fd\mu = \sum_{i=1}^{n} \mu(A_i) y_i$. We also have $\|f\| = \sum_{i=1}^{n} \|y_i\| \chi_{A_i}$ and $\int \|f\| d\mu = \sum_{i=1}^{n} \mu(A_i) \|y_i\|$. It follows that

$$\left| \int fd\mu \right| = \left| \sum_{i=1}^{n} \mu(A_i) y_i \right| \leq \sum_{i=1}^{n} \|\mu(A_i) y_i\| = \sum_{i=1}^{n} \mu(A_i) \|y_i\| = \int \|f\| d\mu.$$

We also have

$$\int \|f\| d\mu = \sum_{i=1}^{n} \mu(A_i) \|y_i\| \leq \mu(\{x \in X \mid f(x) \neq 0\}) \max_{1 \leq i \leq n} \|y_i\|$$

$$= \mu(\{x \in X \mid f(x) \neq 0\}) \|f\|_\infty.$$

(6) Since by (1) the integral is linear, we have

$$N_1(\lambda f) = \int \|\lambda f\| d\mu = \int |\lambda| \|f\| d\mu = |\lambda| \int \|f\| d\mu = |\lambda| N_1(f).$$

Since $\|(f + g)(x)\| \leq \|f(x)\| + \|g(x)\|$ for all $x \in X$, by (5) we have

$$N_1(f + g) = \int \|f + g\| d\mu \leq \int \|f\| d\mu + \int \|g\| d\mu = N_1(f) + N_1(g).$$
Assume that $N_1(f) = 0$, which means that $\int \|f\| \, d\mu = 0$. Since $f$ is a $\mu$-step function, we can write

$$f = \sum_{i=1}^{n} y_i \chi_{A_i},$$

for a finite sequence $(A_1, \ldots, A_n)$ of nonempty pairwise disjoint subsets $A_i \in \mathcal{A}$ of finite measure. Since

$$\|f\| = \sum_{i=1}^{n} \|y_i\| \chi_{A_i},$$

so

$$N_1(f) = \int \|f\| \, d\mu = \sum_{i=1}^{n} \|y_i\| \mu(A_i) = 0.$$

Since $\|y_i\| \geq 0$ and $\mu(A_i) \geq 0$, the following must hold:

- if $\mu(A_i) \neq 0$, then $\|y_i\| = 0$, that is, $y_i = 0$.
- if $y_i \neq 0$, that is, $\|y_i\| \neq 0$, then $\mu(A_i) = 0$.

Consequently

$$\{x \in X \mid f(x) \neq 0\} = \bigcup_{i \in I} A_i, \quad \text{with} \quad I = \{i \mid 1 \leq i \leq n \mid y_i \neq 0\},$$

where $\bigcup_{i \in I} A_i \in \mathcal{A}$ is a set of measure 0, since $i \in I$ implies that $\mu(A_i) = 0$. 

By Proposition 4.15(6), the set

$$\mathcal{SN} = \{f \in \text{Step}_\mu(X, \mathcal{A}, F) \mid N_1(f) = 0\} = \{f \in \text{Step}_\mu(X, \mathcal{A}, F) \mid f = 0 \text{ a.e.}\}$$

is a subspace of $\text{Step}_\mu(X, \mathcal{A}, F)$.

**Definition 4.8.** Let $\text{Step}_\mu(X, \mathcal{A}, F)$ be the quotient space $\text{Step}_\mu(X, \mathcal{A}, F) / \mathcal{SN}$.

For every equivalence class $f \in \text{Step}_\mu(X, \mathcal{A}, F)$, we can define

$$\int f \, d\mu = \int f \, d\mu$$

for any function $f \in \text{Step}_\mu(X, \mathcal{A}, F)$ in the equivalence class of $f$, because if $f = g$ a.e., then $\int f \, d\mu = \int g \, d\mu$, so $\int f \, d\mu$ does not depend on the representative chosen in the equivalence class $f$. Similarly, we define $N_1(f)$ by

$$N_1(f) = N_1(f) = \int \|f\| \, d\mu,$$

for any function $f \in \text{Step}_\mu(X, \mathcal{A}, F)$ in the equivalence class of $f$. Again if $f = g$ a.e., then $\|f\| = \|g\|$ a.e., so $N_1(f) = N_1(g)$, which means that $N_1(f)$ is well defined. It is immediately
verified that $N_1$ is a semi-norm, and in fact a norm, since $N_1(f) = 0$ iff $N_1(f) = 0$ for any representative $f \in \text{Step}_\mu(X, \mathcal{A}, F)$ in the equivalence class $f$ iff $f = 0$ a.e., which means that $f = 0$. Therefore, $(\text{Step}_\mu(X, \mathcal{A}, F), N_1)$ is a normed vector space. It is easy to see that the inequality
\[ \left\| \int f \, d\mu \right\| \leq \int \|f\| \, d\mu = N_1(f) \]
holds, which shows that the map $\int : \text{Step}_\mu(X, \mathcal{A}, F) \to F$ is continuous (in fact, uniformly continuous). The space $(\text{Step}_\mu(X, \mathcal{A}, F), N_1)$ is not complete, so we can apply Theorem 13.68) to form its completion $L_\mu(X, \mathcal{A}, F)$ and extend the map $\int$ to it. Theoretically we have achieved our goal of defining a notion of integral on a normed vector space $L_\mu(X, \mathcal{A}, F)$ which is complete and in which $\text{Step}_\mu(X, \mathcal{A}, F)$ is dense, but the elements in this abstract completion are equivalence classes of Cauchy sequences, and are not easily identifiable with functions.

We will follow a different path, still very much inspired by the completion method involving Cauchy sequences, the twist being that we consider Cauchy sequences whose limit is known ahead of time, but where we use pointwise convergence almost everywhere, instead of pointwise convergence.

### 4.5 Integrable Functions; the Spaces $L_\mu(X, \mathcal{A}, F)$ and $L_\mu(X, \mathcal{A}, F)$

In this section we construct the completion $L_\mu(X, \mathcal{A}, F)$ of the vector space $\text{Step}_\mu(X, \mathcal{A}, F)$ equipped with the semi-norm $N_1$, and construct the integral of a function in $L_\mu(X, \mathcal{A}, F)$. The semi-norm $N_1$ is extended to $L_\mu(X, \mathcal{A}, F)$ as a semi-norm $\|\|_1$ called the $L^1$-semi-norm, and we find that the space of functions such that $\|f\|_1 = 0$ is the set $\mathcal{N}$ of functions in $L_\mu(X, \mathcal{A}, F)$ that are zero a.e. Then we define the quotient space $L_\mu(X, \mathcal{A}, F) = L_\mu(X, \mathcal{A}, F)/\mathcal{N}$. The space $L_\mu(X, \mathcal{A}, F)$ is the completion of $\text{Step}_\mu(X, \mathcal{A}, F)$; this is one of the most important results of this section (the Fischer–Riesz theorem).

As in the previous section the results that we state without proof are proved either in Marle [27] or in Lang [24].

Recall the following definitions.

**Definition 4.9.** A sequence $(f_n)$ of functions $f_n \in \text{Step}_\mu(X, \mathcal{A}, F)$ is a $N_1$-Cauchy sequence if for every $\epsilon > 0$, there is some $N > 0$, such that for all $m, n \geq N$, we have $N_1(f_m - f_n) < \epsilon$, where $N_1(f_m - f_n) = \int \|f_m - f_n\| \, d\mu$. A sequence $(f_n)$ of maps $f_n \in \text{Step}_\mu(X, \mathcal{A}, F)$ converges pointwise almost everywhere to a limit $f : X \to F$ if there is a null set $Z$ such that for every $x \in X - Z$, for every $\epsilon > 0$, there is some $N > 0$, such that $\|f(x) - f_n(x)\| < \epsilon$ for all $n \geq N$.

We define the space $L_\mu(X, \mathcal{A}, F)$ as follows.
Definition 4.10. Let \((X, \mathcal{A}, \mu)\) be a measure space and let \((F, \mathcal{B})\) be a measurable space, with \(F\) a Banach space and \(\mathcal{B}\) its Borel \(\sigma\)-algebra. The set \(L_\mu(X, \mathcal{A}, F)\) of \(\mu\)-integrable functions consists of all functions \(f : X \to F\) such that there is some \(N_1\)-Cauchy sequence \((f_n)_{n \geq 1}\) of \(\mu\)-step maps \(f_n \in \text{Step}_\mu(X, \mathcal{A}, F)\) which converges pointwise almost everywhere to \(f\). A sequence \((f_n)_{n \geq 1}\) of \(\mu\)-step maps as above is called an approximation sequence for \(f\).

Observe that not only do we require that the sequence \((f_n)_{n \geq 1}\) converges pointwise to \(f\) a.e., which makes \(f\) a \(\mu\)-measurable map, but also that this sequence is \(N_1\)-Cauchy. This is the key to defining the notion of integral of the function \(f\), as shown technically in Proposition 4.16.

We will see that \(L_\mu(X, \mathcal{A}, F)\) is a vector space containing \(\text{Step}_\mu(X, \mathcal{A}, F)\), and a subspace of \(M_\mu(X, \mathcal{A}, F)\). Also, and this is the point of the construction, \(L_\mu(X, \mathcal{A}, F)\) is complete with respect to the extension \(\|\|_1\) of the semi-norm \(N_1\) to \(L_\mu(X, \mathcal{A}, F)\), a fact that is not obvious at all from the definition.

The crucial point is that Definition 4.10 is designed so that the following fact holds.

Proposition 4.16. For any \(N_1\)-Cauchy sequence \((f_n)_{n \geq 1}\) of \(\mu\)-step maps, the sequence of integrals \((\int f_n d\mu)_{n \geq 1}\) is a Cauchy sequence in \(F\).

Proof. Indeed, by Proposition 4.15(3), we have

\[
\left\| \int f_n d\mu - \int f_m d\mu \right\| = \left\| \int (f_n - f_m) d\mu \right\| \leq \int \| f_n - f_m \| d\mu = N_1(f_n - f_m),
\]

and since by hypothesis \((f_n)\) is an \(N_1\)-Cauchy sequence, the sequence \((\int f_n d\mu)_{n \geq 1}\) is a Cauchy sequence in \(F\). Indeed, for every \(\epsilon > 0\), since the sequence \((f_n)\) is \(N_1\)-Cauchy, there is some \(N > 0\) such that \(N_1(f_n - f_m) < \epsilon\) for all \(m, n \geq N\), which implies that \(\left\| \int f_n d\mu - \int f_m d\mu \right\| < \epsilon\) for all \(m, n \geq N\).

Then, since \(F\) is complete, the sequence \((\int f_n d\mu)_{n \geq 1}\) converges to an element of \(f\), and if \((f_n)_{n \geq 1}\) is an approximation sequence for \(f \in L_\mu(X, \mathcal{A}, F)\), it is natural to define the integral of \(f\) as

\[
\int f d\mu = \lim_{n \to \infty} \int f_n d\mu.
\]

The problem is that the definition of \(\int f d\mu\) depends on the approximation sequence \((f_n)_{n \geq 1}\) chosen for \(f\).

Actually, the definition of \(\int f d\mu\) does not depend on the approximation approximation sequence \((f_n)_{n \geq 1}\) chosen for \(f\), but proving this is nontrivial. The proof relies on a remarkable fact called the fundamental lemma of integration by Serge Lang; see [24], Chapter VI, §3.
Proposition 4.17. Let \((f_n)_{n \geq 1}\) be any \(N_1\)-Cauchy sequence of maps \(f_n \in \text{Step}_\mu(X, \mathcal{A}, F)\). There exists a subsequence \((f_{i_n})\) which converges pointwise almost everywhere to a limit \(f : X \to F\). Furthermore, for any \(\epsilon > 0\), there is a measurable subset \(Z \in \mathcal{A}\) such that \(\mu(Z) \leq \epsilon\), and the subsequence \((f_{i_n})\) converges uniformly to \(f\) on \(X - Z\) (recall Definition 1.5).

The proof of Proposition 4.17 can be found in Lang [24] (Chapter VI, §3, Lemma 3.1). It should be mentioned that in general, the original sequence \((f_n)\) may not converge pointwise, even a.e. An example of such a sequence \((f_n)\) which is \(N_1\)-Cauchy, yet \((f_n(x))\) diverges for every \(x \in X\), is given in Schwartz [41] (Chapter 5, §6).

Using Proposition 4.17, the following result can be shown. This result implies that the integral \(\int f d\mu\) is well defined.

Proposition 4.18. Let \((f_n)_{n \geq 1}\) and \((g_n)_{n \geq 1}\) be two \(N_1\)-Cauchy sequences of \(\mu\)-step maps \(f_n, g_n \in \text{Step}_\mu(X, \mathcal{A}, F)\) which approximate the same function \(f\). The sequences \(\left(\int f_n d\mu\right)_{n \geq 1}\) and \(\left(\int g_n d\mu\right)_{n \geq 1}\) converges to the same limit, and

\[
\lim_{n \to \infty} \int \|f_n - g_n\| d\mu = 0,
\]

that is, \(\lim_{n \to \infty} N_1(f_n - g_n) = 0\).

The proof of Proposition 4.18 can be found in Lang [24] (Chapter VI, §3, Lemma 3.2).

Proposition 4.18 justifies the following definition.

Definition 4.11. Let \((X, \mathcal{A}, \mu)\) be a measure space and let \((F, \mathcal{B})\) be a measurable space, with \(F\) a Banach space and \(\mathcal{B}\) its Borel \(\sigma\)-algebra. For any function \(f \in \mathcal{L}_\mu(X, \mathcal{A}, F)\), we define the integral\(^3\) of \(f\) (with respect to \(\mu\)) by

\[
\int f d\mu = \lim_{n \to \infty} \int f_n d\mu,
\]

where \((f_n)_{n \geq 1}\) is any approximation sequence of \(f\) by \(\mu\)-step maps.

Proposition 4.19. For any function \(f \in \mathcal{L}_\mu(X, \mathcal{A}, F)\) and any approximation sequence \((f_n)\) of \(f\) with \(f_n \in \text{Step}_\mu(X, \mathcal{A}, F)\), we have \(\|f\| \in \mathcal{L}_\mu(X, \mathcal{A}, \mathbb{R})\), and the sequence \((\|f_n\|)\) is an approximation sequence of \(\|f\|\) with \(\|f_n\| \in \text{Step}_\mu(X, \mathcal{A}, \mathbb{R})\). Furthermore,

\[
\int \|f\| d\mu = \lim_{n \to \infty} \int \|f_n\| d\mu = \lim_{n \to \infty} N_1(f_n).
\]

\(^3\)This integral is usually called the Lebesgue integral or Bochner integral.
Proof. Since the sequence \((f_n)\) converges pointwise to \(f\) a.e., we verify immediately that the sequence \((\|f_n\|)\) converges pointwise to \(\|f\|\) a.e. Since
\[
|\|f_n\| - \|f_m\| | \leq \|f_n - f_m\|
\]
(see just after Definition 13.3), by Proposition 4.15(5) we have
\[
N_1(\|f_n\| - \|f_m\|) = \int |\|f_n\| - \|f_m\|| \, d\mu \leq \int \|f_n - f_m\| \, d\mu = N_1(f_n - f_m),
\]
and since \((f_n)\) is an \(N_1\)-Cauchy sequence, the sequence \((\|f_n\|)\) is an \(N_1\)-Cauchy sequence. Therefore \(\|f\| \in L_\mu(X, A, \mathbb{R})\), and \((\|f_n\|)\) is an approximation sequence of \(\|f\|\). By definition of the integral,
\[
\int \|f\| \, d\mu = \lim_{n \to \infty} \int \|f_n\| \, d\mu = \lim_{n \to \infty} N_1(f_n),
\]
as claimed.

Definition 4.12. For any function \(f \in L_\mu(X, A, F)\), we define the \(L^1\)-semi-norm \(\|f\|_1\) of \(f\) as
\[
\|f\|_1 = \int \|f\| \, d\mu.
\]
Observe that if \(f \in \text{Step}_\mu(X, A, F)\), then \(\|f\|_1 = N_1(f)\). The following proposition is easily shown by passing to the limit.

Proposition 4.20. The set \(L_\mu(X, A, F)\) is a vector space, and \(\|\|_1\) is a semi-norm on \(L_\mu(X, A, F)\). The space \(\text{Step}_\mu(X, A, F)\) is a subspace of \(L_\mu(X, A, F)\), which is a subspace of \(M_\mu(X, A, F)\).

We are almost ready to prove that \(L_\mu(X, A, F)\) is complete with respect to the \(L^1\)-semi-norm, but first we need the following result.

Proposition 4.21. The subspace \(\text{Step}_\mu(X, A, F)\) is dense in \(L_\mu(X, A, F)\) with respect to the \(L^1\)-semi-norm \(\|\|_1\). Furthermore, any approximation sequence \((f_n)_{n \geq 1}\) of \(f\) by \(\mu\)-step maps converges to \(f\) according to the semi-norm \(\|\|_1\).

Proof. Pick any \(f \in L_\mu(X, A, \mathbb{R})\) and let \((f_n)\) be any approximation sequence for \(f\). This means that the sequence \((f_n)\) is a \(N_1\)-Cauchy sequence of \(\mu\)-step maps which converges pointwise to \(f\) a.e. We will prove that
\[
\lim_{n \to \infty} \|f - f_n\|_1 = 0,
\]
which shows that the sequence \((f_n)\) converges to \(f\) in the \(L^1\)-semi-norm.
First, we claim that for any fixed \( n \geq 1 \), the sequence \( (\|f_p - f_n\|)_{p \geq 1} \) is an \( N_1 \)-Cauchy sequence which converges to \( \|f - f_n\| \) a.e. Indeed, we have
\[
\int |\|f_p - f_n\| - \|f_q - f_n\|| \, d\mu \leq \int \|f_p - f_n - (f_q - f_n)\| \, d\mu = \int \|f_p - f_q\| \, d\mu = N_1(f_p - f_q),
\]
and since \( (f_n) \) is a \( N_1 \)-Cauchy sequence, for every \( \epsilon > 0 \), there is some \( N > 0 \) such that \( N_1(f_p - f_q) < \epsilon \) for all \( p, q \geq N \), which shows that \( (\|f_p - f_n\|)_{p \geq 1} \) is a \( N_1 \)-Cauchy sequence (in \( \mathbb{R} \)). The fact that \( (f_p)_{p \geq 1} \) converges pointwise a.e. to \( f \) immediately implies that \( \|f - f_n\| \) converges to \( \|f - f_n\| \) a.e. By definition of \( \|\|_1 \) and of the integral
\[
\|f - f_n\|_1 = \int \|f - f_n\| \, d\mu = \lim_{p \to \infty} \int \|f_p - f_n\| \, d\mu = \lim_{p \to \infty} N_1(f_p - f_n).
\]
Thus for every \( \epsilon > 0 \), there is some \( M_1 > 0 \) such that
\[
\|f - f_n\|_1 - N_1(f_p - f_n) < \frac{\epsilon}{2} \quad \text{for all } p \geq M_1,
\]
and since \( (f_n) \) is an \( N_1 \)-Cauchy sequence, there is some \( M_2 > 0 \) such that
\[
N_1(f_p - f_n) < \frac{\epsilon}{2} \quad \text{for all } n, p \geq M_2,
\]
so for all \( n, p \geq \max(M_1, M_2) \) we have
\[
\|f - f_n\|_1 \leq \|f - f_n\|_1 - N_1(f_p - f_n) + N_1(f_p - f_n) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,
\]
which proves that \( \lim_{n \to \infty} \|f - f_n\|_1 = 0 \); that is, the sequence \( (f_n) \) converges to \( f \) in the \( L^1 \)-semi-norm. \( \square \)

**Remark:** It appears that Lang [24] skipped this step, which is used in the proof his Theorem 3.4, and the proof of the next theorem.

Now we can prove one of our main theorems.

**Theorem 4.22.** (*Fischer–Riesz*) The space \( \mathcal{L}_\mu(X, A, F) \) is complete with respect to the \( L^1 \)-semi-norm. This means that for every sequence \( (f_n)_{n \geq 1} \) of functions \( f_n \in \mathcal{L}_\mu(X, A, F) \), if \( (f_n) \) is \( \|\|_1 \)-Cauchy, then there is some function \( f \in \mathcal{L}_\mu(X, A, F) \) such that for every \( \epsilon > 0 \), there is some \( N > 0 \) such that \( \|f - f_n\|_1 < \epsilon \) for all \( n \geq N \).

**Proof.** Let \( (f_n)_{n \geq 1} \) be an \( \|\|_1 \)-Cauchy sequence of functions \( f_n \in \mathcal{L}_\mu(X, A, F) \). By Proposition 4.21, for every \( n \) there is an approximation sequence \( (g_{n,m})_{m \geq 1} \) of \( \mu \)-step map that
converges to \( f_n \) pointwise a.e., and in the \( \| \cdot \|_1 \)-semi-norm. Thus, for every \( n \geq 1 \), there is some \( m(n) \) such that
\[
\| f_n - g_{n,m(n)} \|_1 \leq \frac{1}{n}.
\]
Each sequence \((g_{n,m(n)})_{n \geq 1}\) is \( N_1 \)-Cauchy, because
\[
N_1(g_{p,m(p)} - g_{q,m(q)}) = \| g_{p,m(p)} - g_{q,m(q)} \|_1 \\
\leq \| g_{p,m(p)} - f_p \|_1 + \| f_p - f_q \|_1 + \| f_q - g_{q,m(q)} \|_1 \\
\leq \frac{1}{p} + \frac{1}{q} + \| f_p - f_q \|_1,
\]
and the right-hand side tends to 0 when \( p \) and \( q \) tend to \(+\infty\), since the sequence \((f_n)\) is \( \| \cdot \|_1 \)-Cauchy. By Proposition 4.17, for each sequence \((g_{n,m(n)})_{m \geq 1}\), we can extract a subsequence \((g_{n_k,m(n_k)})_{k \geq 1}\) that converges pointwise a.e. to some function \( f \in \mathcal{L}_\mu(X, \mathcal{A}, F) \), and is also \( N_1 \)-Cauchy. By the second part of Proposition 4.21, the subsequence \((g_{n_k,m(n_k)})_{k \geq 1}\) converges to \( f \) for the semi-norm \( \| \cdot \|_1 \). Since \((g_{n,m(n)})_{n \geq 1}\) is \( N_1 \)-Cauchy and has a subsequence \((g_{n_k,m(n_k)})_{k \geq 1}\) \( \| \cdot \|_1 \)-convergent to \( f \), it also \( \| \cdot \|_1 \)-converges to the function \( f \). Using the inequality
\[
\| f - f_n \|_1 \leq \| f - g_{n,m(n)} \|_1 + \| g_{n,m(n)} - f_n \|_1 \\
\leq \| f - g_{n,m(n)} \|_1 + \frac{1}{n},
\]
and since the sequence \((g_{n,m(n)})_{n \geq 1}\) \( \| \cdot \|_1 \)-converges to the function \( f \), we deduce that the sequence \((f_n)_{n \geq 1}\) converges to \( f \) for the semi-norm \( \| \cdot \|_1 \).

In the diagram below, the original sequence \((f_n)_{n \geq 1}\) is shown as the top horizontal row. Below each \( f_n \), we have the approximation sequence \((g_{n,m(n)})_{m \geq 1}\) shown as an ascending column. The sequence of \( g_{n,m(n)} \) chosen for each \( n \) is shown in boldface, and its subsequence in red.

\[
\begin{array}{cccccccc}
  f_1 & f_2 & f_3 & f_4 & f_5 & f_6 & \ldots & f_n & \ldots \\
  \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots \\
  g_{1,m(n)} & g_{2,m(n)} & g_{3,m(n)} & g_{4,m(n)} & g_{5,m(n)} & g_{6,m(n)} & \ldots & g_{n,m(n)} & \ldots \\
  \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots \\
  g_{1,6} & g_{2,6} & g_{3,6} & g_{4,6} & g_{5,6} & g_{6,6} & \ldots & g_{n,6} & \ldots \\
  g_{1,5} & g_{2,5} & g_{3,5} & g_{4,5} & g_{5,5} & g_{6,5} & \ldots & g_{n,5} & \ldots \\
  g_{1,4} & g_{2,4} & g_{3,4} & g_{4,4} & g_{5,4} & g_{6,4} & \ldots & g_{n,4} & \ldots \\
  g_{1,3} & g_{2,3} & g_{3,3} & g_{4,3} & g_{5,3} & g_{6,3} & \ldots & g_{n,3} & \ldots \\
  g_{1,2} & g_{2,2} & g_{3,2} & g_{4,2} & g_{5,2} & g_{6,2} & \ldots & g_{n,2} & \ldots \\
  g_{1,1} & g_{2,1} & g_{3,1} & g_{4,1} & g_{5,1} & g_{6,1} & \ldots & g_{n,1} & \ldots \\
\end{array}
\]

This concludes the proof. \( \square \)

The following properties of the integral are easily obtained by passing to the limit.
Proposition 4.23. Let \((X, \mathcal{A}, \mu)\) be a measure space and let \((F, \mathcal{B})\) be a measurable space, with \(F\) a Banach space and \(\mathcal{B}\) its Borel \(\sigma\)-algebra. The following properties hold:

1. For any \(f \in \mathcal{L}_\mu(X, \mathcal{A}, F)\), if \(f = 0\) a.e., then \(\int fd\mu = 0\). More generally, if \(f, g \in \mathcal{L}_\mu(X, \mathcal{A}, F)\) and \(f = g\) a.e., then \(\int fd\mu = \int gd\mu\).

2. For any \(f \in \mathcal{L}_\mu(X, \mathcal{A}, F)\), and for any measurable subset \(A \in \mathcal{A}\), the integral \(\int_A fd\mu = \int f \chi_A d\mu\) exists, and
\[
\left\| \int_A fd\mu \right\| \leq \int_A \|f\| d\mu \leq \|f\|_\infty \mu(A).
\]
Furthermore, if \(A, B \in \mathcal{A}\) are disjoint, then
\[
\int_{A \cup B} fd\mu = \int_A fd\mu + \int_B fd\mu.
\]

3. The integral \(\int : \mathcal{L}_\mu(X, \mathcal{A}, F) \to F\) is linear.

4. For any \(f \in \mathcal{L}_\mu(X, \mathcal{A}, F)\), we have \(\|f\| \in \mathcal{L}_\mu(X, \mathcal{A}, \mathbb{R})\), and
\[
\left\| \int fd\mu \right\| \leq \int \|f\| d\mu = \|f\|_1.
\]

5. If \(f, g \in \mathcal{L}_\mu(X, \mathcal{A}, \mathbb{R})\), then \(\sup(f, g), \inf(f, g), f^+, f^-, |f| \in \mathcal{L}_\mu(X, \mathcal{A}, \mathbb{R})\). Since \(f^+ = (|f| + f)/2\) and \(f^- = (|f| - f)/2\), we have \(f \in \mathcal{L}_\mu(X, \mathcal{A}, \mathbb{R})\) iff \(f^+ \in \mathcal{L}_\mu(X, \mathcal{A}, \mathbb{R})\) and \(f^- \in \mathcal{L}_\mu(X, \mathcal{A}, \mathbb{R})\).

6. If \(f, g \in \mathcal{L}_\mu(X, \mathcal{A}, \mathbb{R})\) and \(f \leq g\) a.e., then \(\int fd\mu \leq \int gd\mu\). In particular, if \(f \geq 0\) a.e., then \(\int fd\mu \geq 0\).

7. Let \(F_1\) and \(F_2\) be two Banach spaces, and let \(h : F_1 \to F_2\) be a linear (or semi-linear map when the field is \(\mathbb{C}\)). If \(f \in \mathcal{L}_\mu(X, \mathcal{A}, F_1)\), then \(h \circ f \in \mathcal{L}_\mu(X, \mathcal{A}, F_2)\), and
\[
\int (h \circ f) d\mu = h \left( \int fd\mu \right).
\]

8. Let \(F_1\) and \(F_2\) be two Banach spaces, and let \(F_1 \times F_2\) be the product space (under any of the product norms defined just before Definition 13.13). Then there is an isomorphism between \(\mathcal{L}_\mu(X, \mathcal{A}, F_1 \times F_2)\) and \(\mathcal{L}_\mu(X, \mathcal{A}, F_1) \times \mathcal{L}_\mu(X, \mathcal{A}, F_2)\), and if \(f = (f_1, f_2)\), then
\[
\int fd\mu = \left( \int f_1 d\mu, \int f_2 d\mu \right).
\]
In particular, since \(\mathbb{C}\) is isomorphic to \(\mathbb{R} \times \mathbb{R}\), a function \(f \in \mathcal{L}_\mu(X, \mathcal{A}, \mathbb{C})\) corresponds uniquely to a function \(f = u + iv\) with \(u, v \in \mathcal{L}_\mu(X, \mathcal{A}, \mathbb{R})\), and we have
\[
\int fd\mu = \int ud\mu + i \int vd\mu.
\]
Remark: Observe that in our approach, if \( f \) is a real-valued function or a complex-valued function, the integral \( \int f \, d\mu \) is defined directly. There is another approach in which the integral is first defined for real-valued positive functions. Then the integral of a real-valued function \( f \) is defined in terms of the integrals of \( f^+ \) and \( f^- \), and the integral of a complex valued function \( f = u + iv \) is defined in terms of the integrals of \( u^+, u^-, v^+, v^- \). See Rudin [36], Definition 1.31.

The next step is to identify the functions \( f \) in \( \mathcal{L}_\mu(X, \mathcal{A}, \mathbb{R}) \) such that \( \|f\|_1 = 0 \). For this, we need two propositions.

**Proposition 4.24.** For any function \( f \in \mathcal{L}_\mu(X, \mathcal{A}, F) \), and for any real \( a > 0 \), the subset \( E_a = \{ x \in X \mid \|f(x)\| \geq a \} \) can be written as \( E_a = (B - Z) \cup N \), with \( B \) a measurable subset of finite measure, and \( Z \) and \( N \) two null subsets. The function \( f \) vanishes outside of a \( \sigma \)-finite measurable set.

**Proof.** By Proposition 4.17, there is an \( N_1 \)-Cauchy sequence \( (f_n) \) of \( \mu \)-step maps which converges pointwise to \( f \) a.e., and for every \( \epsilon > 0 \), there is a measurable subset \( Z_1 \) of measure \( \mu(Z_1) < \epsilon \) such that \( f_n \) converges uniformly to \( f \) on \( X - Z_1 \). Pick \( \epsilon' \) such that \( 0 < \epsilon' < a \). The uniform convergence implies that there is some \( M > 0 \) such that for all \( n \geq M \) and all \( x \in X - Z_1 \),

\[
\|f(x) - f_n(x)\| \leq \epsilon',
\]

thus

\[
\|f(x)\| \leq \|f_n(x)\| + \epsilon',
\]

which implies that

\[
E_a \subseteq Z_1 \cup \{ x \in X \mid \|f_n(x)\| \geq a - \epsilon' \},
\]

where both sets on the right-hand side have finite measure (the second one because \( f_n \) is a \( \mu \)-step function, and so is \( \|f_n\| \), and a \( \mu \)-step function vanishes outside of a set of finite measure). Since the function \( f \) is \( \mu \)-measurable, by Proposition 4.12(1), it is equal a.e. to a measurable function \( g \), so there is a null set such \( Z \) that \( \|f\|(x) = g(x) \) for all \( x \in X - Z \). Then we have

\[
E_a = \{ x \in X \mid \|f(x)\| \geq a \}
\]

\[
= \{ x \in X - Z \mid \|f(x)\| \geq a \} \cup \{ x \in Z \mid \|f(x)\| \geq a \}
\]

\[
= \{ x \in X - Z \mid g(x) \geq a \} \cup \{ x \in Z \mid \|f(x)\| \geq a \}
\]

\[
= (\{ x \in X \mid g(x) \geq a \} - Z) \cup \{ x \in Z \mid \|f(x)\| \geq a \}
\]

\[
= (B - Z) \cup N,
\]

with \( B = \{ x \in X \mid g(x) \geq a \} \) and \( N = \{ x \in Z \mid \|f(x)\| \geq a \} \). Since \( g \) is measurable and \( [a, \infty) \) is closed, \( B \) is measurable, and \( N \) as a subset of a null set is a null set.

What we showed above with \( \|f\| \) replaced by \( g \) measurable implies that \( B \) has finite measure. The second statement of the proposition follows from Proposition 4.12(2). \( \square \)
Definition 4.13. Let which is just the subspace of function equal to 0 a.e. Then we define \( L_\mu \)

Proposition 4.25. Let \((f_n)_{n \geq 1}\) be any \(\| \|_1\)-Cauchy sequence of maps \(f_n \in L_\mu(X,A,F)\) that converges to some function \(f \in L_\mu(X,A,F)\) in the semi-norm \(\| \|_1\). There exists a subsequence \((f_{n_k})\) which converges pointwise almost everywhere to \(f\). Furthermore, for any \(\epsilon > 0\), there is a measurable subset \(Z \subset A\) such that \(\mu(Z) \leq \epsilon\), and the subsequence \((f_{n_k})\) converges uniformly to \(f\) on \(X - Z\) (recall Definition 1.5).

Here are some corollaries of Proposition 4.25.

Proposition 4.26. For any function \(f \in L_\mu(X,A,F)\), we have \(\|f\|_1 = 0\) iff \(f = 0\) a.e.

Proof. If \(f = 0\) a.e., then \(\|f\| = 0\) a.e., and by Proposition 4.23(1), we have \(\|f\|_1 = 0\). Conversely, the sequence \((f_n)\) where \(f_n\) is the zero function is \(\| \|_1\)-Cauchy and converges to \(f\) in the \(\| \|_1\)-norm. By Proposition 4.25 there is a subsequence that converges pointwise a.e. to \(f\). But since \(f_n\) is the zero function for all \(n\), this subsequence also converges pointwise a.e. to the zero function, so \(f = 0\) a.e.

Proposition 4.26 is the second important result of this section because it provides a very natural characterization of the functions \(f\) such that \(\|f\|_1 = 0\).

Proposition 4.27. Let \((f_n)\) be a sequence of functions \(f_n \in L_\mu(X,A,F)\). If \((f_n)\) is an \(\| \|_1\)-Cauchy sequence which converges pointwise a.e. to a function \(f: X \to F\), then \(f \in L_\mu(X,A,F)\), and \((f_n)\) converges to \(f\) in the semi-norm \(\| \|_1\).

Proof. Since the sequence \((f_n)\) is an \(\| \|_1\)-Cauchy sequence, by the Fischer–Riesz theorem (Theorem 4.22) it converges to some function \(g \in L_\mu(X,A,F)\) in the \(\| \|_1\)-semi-norm. By Proposition 4.25, some subsequence \((f_{n_k})_{k \geq 1}\) of \((f_n)\) converges pointwise a.e. to \(g\). Since \((f_n)\) converges pointwise a.e. to \(f\), the subsequence \((f_{n_k})_{k \geq 1}\) also converges pointwise a.e. to \(f\), so \(f = g\) a.e., and since \(g \in L_\mu(X,A,F)\) and \((f_n)\) converges to \(g\) in the semi-norm \(\| \|_1\), we also have \(f \in L_\mu(X,A,F)\), and \((f_n)\) converges to \(f\) in the semi-norm \(\| \|_1\).

The main disadvantage of the space \(L_\mu(X,A,F)\) is that it is not a normed vector space under the semi-norm \(\| \|_1\). Thus it is natural to consider the quotient of \(L_\mu(X,A,F)\) by the subspace \(N\) consisting of the functions such that \(\|f\|_1 = 0\).

Definition 4.13. Let \(N\) be the subspace of \(L_\mu(X,A,F)\) given by

\[
N = \{ f \in L_\mu(X,A,F) \mid \|f\|_1 = 0 \},
\]

which is just the subspace of function equal to 0 a.e. Then we define \(L_\mu(X,A,F)\) as the quotient space

\[
L_\mu(X,A,F) = L_\mu(X,A,F)/N.
\]
For any equivalence class \( f \in \mathcal{L}_\mu(X,A,F) \), since for any two representatives \( f,g \in \mathcal{L}_\mu(X,A,F) \) in the equivalence class \( f \), we have \( f = g \) a.e., by Proposition 4.23(1),

\[
\int f \, d\mu = \int g \, d\mu,
\]

so we can define \( \int f \, d\mu \) as

\[
\int f \, d\mu = \int f \, d\mu.
\]

Similarly, \( \|f\|_1 \) is defined as

\[
\|f\|_1 = \|f\|_1,
\]

for any \( f \in \mathcal{L}_\mu(X,A,F) \) in the equivalence class \( f \).

The following theorem is immediately obtained from Theorem 4.22 by passing to the quotient.

**Theorem 4.28.** (Fischer–Riesz) The semi-norm \( \|\| \) on \( \mathcal{L}_\mu(X,A,F) \) induced by the semi-norm \( \|\|_1 \) on \( \mathcal{L}_\mu(X,A,F) \) by passing to the quotient is a norm on \( \mathcal{L}_\mu(X,A,F) \) called the \( L^1 \)-norm. With this norm, the space \( \mathcal{L}_\mu(X,A,F) \) is complete (it is a Banach space). The subspace \( \text{Step}_\mu(X,A,F) \) is dense in \( \mathcal{L}_\mu(X,A,F) \).

Finally, the following proposition confirms one of our earlier claims.

**Proposition 4.29.** The space \( \mathcal{L}_\mu(X,A,F) = \mathcal{L}_\mu(X,A,F)/N \) is isomorphic to the Cauchy completion of the space \( \text{Step}_\mu(X,A,F)/SN \); see the diagram

\[
\text{Step}_\mu(X,A,F) \xrightarrow{\text{completion}} \mathcal{L}_\mu(X,A,F) \xrightarrow{\text{quotient}} \mathcal{L}_\mu(X,A,F) = \mathcal{L}_\mu(X,A,F)/N.
\]

In the next section, we consider some fundamental convergence theorems. A very useful corollary of these theorems is that a function \( f \) belongs to \( \mathcal{L}_\mu(X,A,F) \) iff it belongs to \( \mathcal{M}_\mu(X,A,F) \) (it is \( \mu \)-measurable), and if \( \int \|f\| \, d\mu \) exists. By Proposition 4.20, the space \( \mathcal{L}_\mu(X,A,F) \) is a subspace of \( \mathcal{M}_\mu(X,A,F) \), and we already know from Proposition 4.23(4) that if \( f \in \mathcal{L}_\mu(X,A,F) \) then \( \|f\| \in \mathcal{L}_\mu(X,A,\mathbb{R}) \). The converse is not trivial, but it will be shown as a corollary of the dominated convergence theorem discussed in Section 4.6.
4.6 Fundamental Convergence Theorems

Besides the fact that the Lebesgue–Bochner integral is defined for a much bigger class of functions than the regulated functions (or the Riemann-integrable functions), one of its main advantages is that it leads to simple and flexible criteria to tell whether the limit of a sequence of integrable functions is integrable. Most of these results allow interchanging a limit and an integral. We begin with criteria applying to real-valued functions. These results actually apply to extended functions with values in $\mathbb{R} \cup \{+\infty\}$, but for simplicity we stick to functions $f : X \to \mathbb{R}$. As in the previous section the results that we state without proof are proven either in Marle [27] or in Lang [24].

Theorem 4.30. (Monotone Convergence Theorem) Let $(f_n)_{n \geq 1}$ be a sequence of functions $f_n \in \mathcal{L}_\mu(X, \mathcal{A}, \mathbb{R})$ such that $f_n \leq f_{n+1}$ for all $n \geq 1$, and assume that there is some $M > 0$ such that

$$\left| \int f_n d\mu \right| \leq M \quad \text{for all } n \geq 1.$$ 

Then the sequence $(f_n)_{n \geq 1}$ converges pointwise a.e., and also in the $\| \cdot \|_1$-norm, to a function $f \in \mathcal{L}_\mu(X, \mathcal{A}, \mathbb{R})$. We also have $\lim_{n \to \infty} \|f_n\|_1 = \|f\|_1$ and

$$\lim_{n \to \infty} \int f_n d\mu = \int f d\mu.$$ 

The same result applies to a nonincreasing sequence $(f_n)$ (with $f_n \geq f_{n+1}$ for all $n \geq 1$).

Proof. We follow Lang [24] (Chapter VI, §5, Theorem 5.5). Let

$$\alpha = \sup_k \int f_k d\mu,$$

which is well defined since

$$\left| \int f_n d\mu \right| \leq M \quad \text{for all } n \geq 1.$$ 

For $n \geq m$ we have

$$\|f_n - f_m\|_1 = \int (f_n - f_m) d\mu$$

$$= \int f_n d\mu - \int f_m d\mu$$

$$\leq \alpha - \int f_m d\mu,$$

which implies that $(f_n)$ is a $\| \cdot \|_1$-Cauchy sequence. By the Fischer–Riesz theorem (Theorem 4.22), the sequence converges to some limit $f \in \mathcal{L}_\mu(X, \mathcal{A}, \mathbb{R})$ in the $\| \cdot \|_1$-norm. By Proposition 4.25, there is a subsequence $(f_{n_k})_{k \geq 1}$ of $(f_n)$ that converges a.e. to $f \in \mathcal{L}_\mu(X, \mathcal{A}, \mathbb{R})$, and since the sequence $(f_n)$ is increasing, it also converges a.e. to $f$. \qed
The following theorem has a different flavor. It asserts the existence of the sup of a sequence of functions.

**Theorem 4.31.** (Beppo–Levi) Let \((f_n)\) be a sequence of functions \(f_n \in L_\mu(X, A, \mathbb{R})\). If there is a function \(g \in L_\mu(X, A, \mathbb{R})\) such that \(g \geq 0\) and \(|f_n| \leq g\) for all \(n \geq 1\), then \(\sup_{n \geq 1} f_n\) and \(\inf_{n \geq 1} f_n\) belong to \(L_\mu(X, A, \mathbb{R})\), and we have

\[
\sup_{n \geq 1} \int f_n \, d\mu \leq \int (\sup_{n \geq 1} f_n) \, d\mu \quad \text{and} \quad \int (\inf_{n \geq 1} f_n) \, d\mu \leq \inf_{n \geq 1} \int f_n \, d\mu.
\]

*Proof.* We follow Lang [24] (Chapter VI, §5, Corollary 5.6). The functions

\[
g_n = \sup \{f_1, \ldots, f_n\}
\]

belong to \(L_\mu(X, A, \mathbb{R})\) and form an increasing sequence bounded by \(g\). Since

\[
\int g_n \, d\mu \leq \int g_{n+1} \, d\mu \quad \text{and} \quad \int g_n \, d\mu \leq \int g \, d\mu,
\]

there is some \(M > 0\) such that \(|\int g_n \, d\mu| \leq M\) for all \(n \geq 1\). Therefore, by the monotone convergence theorem (Theorem 4.30), the sequence \((g_n)\) converges pointwise a.e. to some function in \(L_\mu(X, A, \mathbb{R})\), but \((g_n)\) converges pointwise to \(\sup_{n \geq 1} f_n\), so the sequence \((g_n)\) converges pointwise a.e. to \(\sup_{n \geq 1} f_n\). Since \(f_n \leq \sup_{n \geq 1} f_n\), we have

\[
\int f_n \, d\mu \leq \int (\sup_{n \geq 1} f_n) \, d\mu,
\]

which implies

\[
\sup_{n \geq 1} \int f_n \, d\mu \leq \int (\sup_{n \geq 1} f_n) \, d\mu,
\]

as claimed. \(\square\)

Given a sequence \((f_n)_{n \geq 1}\) of functions \(f_n: X \rightarrow \mathbb{R}\) such that \(f_n \geq 0\), recall that

\[
\liminf f_n = \lim_{k \to \infty} \inf_{n \geq k} f_n.
\]

**Theorem 4.32.** (Fatou’s Lemma) Let \((f_n)_{n \geq 1}\) be a sequence of functions \(f_n \in L_\mu(X, A, \mathbb{R})\) such that \(f_n \geq 0\). If \(\liminf ||f_n||_1 = \liminf \int f_n \, d\mu\) exists, then there is a function \(f \in L_\mu(X, A, \mathbb{R})\) such that \(\liminf f_n\) converges pointwise to \(f\) a.e., and

\[
\int f \, d\mu \leq \liminf \int f_n \, d\mu.
\]

A proof of Theorem 4.32 is given in Lang [24] (Chapter VI, §5).

The next theorem applies to functions with values in any Banach space \(F\), and is the most important convergence theorem.
Theorem 4.33. (Lebesgue Dominated Convergence Theorem) Let \((f_n)_{n \geq 1}\) be a sequence of functions \(f_n \in L_\mu(X, A, F)\). If \((f_n)\) converges pointwise a.e. to a function \(f : X \to F\), and if there is some function \(g \in L_\mu(X, A, \mathbb{R})\) such that \(g \geq 0\) and \(\|f_n\| \leq g\) for all \(n \geq 1\), then \(f \in L_\mu(X, A, F)\) and \((f_n)_{n \geq 1}\) converges to \(f\) in the \(\|\cdot\|_1\)-norm. Consequently

\[
\lim_{n \to \infty} \int f_n d\mu = \int f d\mu.
\]

Proof. We follow Marle [27] (Chapter 2, Section 4, Theorem 2.4.7). For each \(n, p \geq 1\), let

\[
g_{n, p} = \sup_{n \leq m \leq n + p} \sup_{n \leq r \leq n + p} \|f_m - f_r\|
\]

\[
g_n = \sup_{m, r \geq n} \|f_m - f_r\| = \sup_{p \geq 1} g_{n, p}.
\]

For all \(n, p \geq 1\), we have \(g_{n, p} \in L_\mu(X, A, \mathbb{R})\), \(g_{n, p} \leq g_{n, p+1}\), and

\[
0 \leq g_{n, p} \leq 2g.
\]

We get

\[
\left| \int g_{n, p} d\mu \right| \leq 2 \int g d\mu,
\]

so by the monotone convergence theorem (Theorem 4.30), the sequence \((g_{n, p})_{p \geq 1}\) converges pointwise a.e. to a limit in \(L_\mu(X, A, \mathbb{R})\). However, by construction this limit is \(g_n\). Thus \(g_n \in L_\mu(X, A, \mathbb{R})\), and we also have

\[
\int g_n d\mu \leq 2 \int g d\mu.
\]

The sequence \((g_n)_{n \geq 1}\) is nonincreasing and since by hypothesis \((f_n)\) converges pointwise a.e. to \(f\), the sequence \((g_n)\) converges pointwise a.e. to 0. By the monotone convergence theorem (Theorem 4.30),

\[
\lim_{n \to \infty} \int g_n d\mu = 0.
\]

Hence by definition of \(g_n\) the sequence \((f_n)\) is actually a Cauchy sequence, and by Proposition 4.27, we have \(f \in L_\mu(X, A, F)\) and \((f_n)\) converges to \(f\) in the \(\|\cdot\|_1\)-norm. We have

\[
\left\| \int f_n d\mu - \int f d\mu \right\| = \left\| \int (f_n - f) d\mu \right\|
\]

\[
\leq \int \|f_n - f\| d\mu
\]

\[
= \|f_n - f\|_1,
\]

and since \((f_n)\) converges to \(f\) in the \(\|\cdot\|_1\)-norm, this implies that

\[
\lim_{n \to \infty} \int f_n d\mu = \int f d\mu,
\]

as claimed. \(\square\)
Theorem 4.34. A function $f: X \to F$ is integrable, that is, $f \in \mathcal{L}_\mu(X, \mathcal{A}, F)$, iff $f \in \mathcal{M}_\mu(X, \mathcal{A}, F)$ and $\|f\| \in \mathcal{L}_\mu(X, \mathcal{A}, \mathbb{R})$. More generally, if $f \in \mathcal{M}_\mu(X, \mathcal{A}, F)$ and if there is a function $g \in \mathcal{L}_\mu(X, \mathcal{A}, \mathbb{R})$ such that $g \geq 0$ and $\|f\| \leq g$, then $f \in \mathcal{L}_\mu(X, \mathcal{A}, F)$.

Proof. By Proposition 4.20, the space $\mathcal{L}_\mu(X, \mathcal{A}, F)$ is a subspace of $\mathcal{M}_\mu(X, \mathcal{A}, F)$, and we already know from Proposition 4.23(4) that if $f \in \mathcal{L}_\mu(X, \mathcal{A}, F)$ then $\|f\| \in \mathcal{L}_\mu(X, \mathcal{A}, \mathbb{R})$.

For the converse, we may assume that $f$ and $g$ are measurable, since a $\mu$-measurable function is equal a.e. to a measurable function. There is a sequence $(h_n)_{n \geq 1}$ of $\mu$-step maps that converges pointwise a.e. to $f$. For every $x \in X$ and every $n \geq 1$, let

$$h'_n(x) = \begin{cases} h_n(x) & \text{if } \|h_n(x)\| \leq 2g(x) \\ 0 & \text{if } \|h_n(x)\| > 2g(x). \end{cases}$$

For every $n \geq 1$, the function $h'_n$ is a $\mu$-step function and $\|h'_n\| \leq 2g \in \mathcal{L}_\mu(X, \mathcal{A}, \mathbb{R})$. We claim that for every $x \in X$ such that $(h_n(x))$ converges to $f(x)$, the sequence $(h'_n(x))$ also converges to $f(x)$. If $g(x) = 0$, then $\|f(x)\| = 0$, so $f(x) = 0$, and then $h'_n(x) = 0$ for all $n \geq 1$. If $g(x) \neq 0$, since the sequence $(h_n(x))$ converges to $f(x)$ and since $\|f(x)\| < 2g(x)$, there is some $M > 0$ such that $\|h_n(x)\| \leq 2g(x)$ for all $n \geq M$, which implies that $h'_n(x) = h_n(x)$. It follows that the sequence $(h'_n)_{n \geq 1}$ converges pointwise a.e. to $f$. By Theorem 4.33, since $\|h'_n\| \leq 2g$, we conclude that $f \in \mathcal{L}_\mu(X, \mathcal{A}, F)$.

A useful corollary of Theorem 4.34 is the following result.

Proposition 4.35. The following facts hold:

1. If $f \in \mathcal{L}_\mu(X, \mathcal{A}, F)$, $g \in \mathcal{M}_\mu(X, \mathcal{A}, K)$ with $K = \mathbb{R}$ or $K = \mathbb{C}$, and $\|g\|$ is bounded, then $fg \in \mathcal{L}_\mu(X, \mathcal{A}, F)$.

2. Let $h: E \times F \to G$ be a continuous bilinear map, where $E, F, G$ are Banach spaces. If $f \in \mathcal{L}_\mu(X, \mathcal{A}, E)$ and $g \in \mathcal{M}_\mu(X, \mathcal{A}, F)$ with $\|g\|$ bounded, then $h(f, g) \in \mathcal{L}_\mu(X, \mathcal{A}, G)$.

3. Let $f \in \mathcal{L}_\mu(X, \mathcal{A}, \mathbb{R})$, with $f \geq 0$, and let $gf \in \mathcal{M}_\mu(X, \mathcal{A}, \mathbb{R})$, with values in an interval $[m, M]$. Then $fg \in \mathcal{L}_\mu(X, \mathcal{A}, \mathbb{R})$ and we have

$$m \int f d\mu \leq \int fg d\mu \leq M \int f d\mu.$$

Another corollary involves series of functions in $\mathcal{L}_\mu(X, \mathcal{A}, F)$. 
Proposition 4.36. Let \((f_n)_{n \geq} \) be a sequence of functions \(f_n \in \mathcal{L}_\mu(X, \mathcal{A}, F)\). If the series
\[
\sum_{n=1}^{\infty} \int \|f_n\| \, d\mu
\]
converges, then the series
\[
f(x) = \sum_{n=1}^{\infty} f_n(x)
\]
converges a.e., \(f \in \mathcal{L}_\mu(X, \mathcal{A}, F)\), and
\[
\int f \, d\mu = \sum_{n=1}^{\infty} \int f_n \, d\mu
\]
The following proposition is needed for the proof of several results stated in Chapter 5.

Proposition 4.37. (Averaging Theorem) Let \(f \in \mathcal{L}_\mu(X, \mathcal{A}, F)\) be any function and let \(S\) be any closed subset of \(X\). If for any measurable subset \(A\) of finite measure \(\mu(A) > 0\) we have
\[
\frac{1}{\mu(A)} \int_A f \, d\mu \in S,
\]
and if \(0 \in S\) or if \(X\) is \(\sigma\)-finite, then \(f(x) \in S\) for all almost all \(x \in X\).

Proposition 4.37 is proven in Lang [24] (Chapter VI, Theorem 5.15). By applying Proposition 4.37 to the set \(S = \{0\}\), we obtain the following useful corollary.

Proposition 4.38. For any function \(f \in \mathcal{L}_\mu(X, \mathcal{A}, F)\), if
\[
\int_A f \, d\mu = 0
\]
for every measurable subset \(A\) of finite measure, then \(f = 0\) almost everywhere.

We conclude this section with two results about the continuity and the differentiability of a function defined by an integral.

Proposition 4.39. Let \((X, \mathcal{A}, \mu)\) be a measure space, let \(U\) be metric space, let \(F\) be a Banach space (over \(\mathbb{R}\) or \(\mathbb{C}\)), and let \(f: U \times X \to F\) be a function.

1. (Continuity of the integral) Assume that \(f\) has the following properties:

(a) For every \(u \in U\), the map \(f_{u,-}: X \to F\) given by
\[
f_{u,-}(x) = f(u, x) \quad x \in X,
\]
belongs to \(\mathcal{L}_\mu(X, \mathcal{A}, F)\),
(b) For every \( x \in X \), the map \( f_{-,x}: U \to F \) given by
\[
f_{-,x}(u) = f(u, x) \quad u \in U,
\]
is continuous.

(c) There is some \( g \in L_\mu(X, A, \mathbb{R}) \), \( g \geq 0 \), such that
\[
\|f(u, x)\| \leq g(x) \quad \text{for all } u \in U, \text{ and all } x \in X.
\]
Then the map \( h: U \to F \) given by
\[
h(u) = \int f_{u,-} \, d\mu
\]
is continuous.

2. (Taking a derivative under the integral sign) Suppose \( U \) is an open subset of a Banach space \( G \), and let \( \mathcal{L}(G; F) \) be the space of linear continuous maps from \( G \) to \( F \), with the operator norm (see Definition 13.42). Assume that \( f \) has the following properties:

(d) For every \( u \in U \), the map \( f_{u,-}: X \to F \) given by
\[
f_{u,-}(x) = f(u, x) \quad x \in X,
\]
belongs to \( L_\mu(X, A, F) \).

(e) For every \( x \in X \), the map \( f_{-x}: U \to F \) is differentiable, and let \( Df_{-,x} \) be this derivative (a map from \( U \) to \( \mathcal{L}(G; F) \)).

(f) For every \( u \in U \), the map from \( X \) to \( \mathcal{L}(G; F) \) given by
\[
x \mapsto Df_{-x}(u)
\]
belongs to \( L_\mu(X, A, \mathcal{L}(G; F)) \), and there is some \( g \in L_\mu(X, A, \mathbb{R}) \), \( g \geq 0 \), such that
\[
\|Df_{-x}(u)\| \leq g(x) \quad \text{for all } u \in U, \text{ and all } x \in X.
\]
Then the map \( h: U \to F \) given by
\[
h(u) = \int f_{u,-} \, d\mu
\]
is differentiable in \( U \), and its derivative at \( u \in U \) is given by
\[
Dh_u = \int Df_{-x}(u) \, d\mu.
\]
4.6. **FUNDAMENTAL CONVERGENCE THEOREMS**

More could be said about the applications of the convergence theorems, but we have everything we need.

**Remark:** There is another approach to the definition of the integral that applies only to real and complex-valued functions, presented in various texts such as Rudin [36]. In this approach, positive functions play a central role. This approach relies on the fact that for any measurable function $f : X \to [0, +\infty]$ there is a monotonic sequence $(f_n)$ of positive step functions that converges pointwise to $f$; see Rudin [36] (Chapter 1, Theorem 1.17). The integral of a step function is defined in the usual way. Then, given any measurable function $f : X \to [0, +\infty]$, the integral of $f$ is defined as

$$\int f \, d\mu = \sup_{0 \leq s \leq f} \int s \, d\mu,$$

where $s$ is a step function.

A main difference with the approach we followed is that this definition of the integral allows it to take the value $+\infty$. Of course, later on, in order to define what it means for a measurable complex-valued function $f : X \to \mathbb{C}$ to be integrable, the condition

$$\int |f| \, d\mu < +\infty$$

is required. Thus in this approach, the space $L_\mu(X, A, \mathbb{C})$ is defined as the space of measurable functions such that the positive function $|f|$ has a finite integral.

In the approach that we followed, due to Bochner and Dunford, the space $L_\mu(X, A, \mathbb{C})$ is defined in terms of various Cauchy sequences, and the fact that if a function $f : X \to \mathbb{C}$ is measurable and if $|f|$ has a finite integral, then $f \in L_\mu(X, A, \mathbb{C})$, is a theorem (Theorem 4.34). Ultimately, it is proved that $L_\mu(X, A, \mathbb{C})$ is complete (see Rudin [36] Chapter 3, Theorem 3.11), and it is observed that as a corollary, from a $\| \cdot \|_1$-Cauchy sequence, one can extract a subsequence that converges pointwise a.e. (Rudin [36] Chapter 3, Theorem 3.12). It is also shown that the $\mu$-step functions are dense in $L_\mu(X, A, \mathbb{C})$ (Rudin [36] Chapter 3, Theorem 3.13).

The circle is closed. What we took as a definition of $L_\mu(X, A, \mathbb{C})$ is obtained as a corollary in the other approach, and the two approaches yield the same notion of integrability (the same space $L_\mu(X, A, \mathbb{C})$).

One might argue that the approach relying on the integral of positive functions is simpler, or at least takes less efforts. For one thing, it does not need the refined notion of $\mu$-step maps and $\mu$-mesurable maps. However, our feeling is that the approach we followed provides a better understanding of the structure of $L_\mu(X, A, \mathbb{C})$. Also, it can’t be avoided in one wants to integrate functions with values in an infinite-dimensional vector space.
4.7 The Spaces $L^p_\mu(X, \mathcal{A}, F)$ and $L^p(X, \mathcal{A}, F)$; $p = 1, 2, \infty$

Theorem 4.34 suggests the definition of other families of integrable functions.

Definition 4.14. Let $(X, \mathcal{A}, \mu)$ be a measure space and let $(F, \mathcal{B})$ be a measurable space, with $F$ a Banach space and $\mathcal{B}$ its Borel $\sigma$-algebra. For any $p \geq 1$, the set of functions $L^p_\mu(X, \mathcal{A}, F)$ is the set of functions $f \in M_\mu(X, \mathcal{A}, F)$ such that $\|f\|^p \in L^{\infty}(X, \mathcal{A}, \mathbb{R})$, or equivalently

\[ \int \|f\|^p \, d\mu < +\infty. \]

By Theorem 4.34, we have $L^1_\mu(X, \mathcal{A}, F) = L_\mu(X, \mathcal{A}, F)$, and we know that $L^1_\mu(X, \mathcal{A}, F)$ is a vector space. Although it is possible to develop a theory of $L^p$ spaces for any $p \geq 1$, for our applications to harmonic analysis we only need the cases $p = 1, 2$. The case where $p = \infty$ arises when we consider duality, but we postpone the definition of $L^\infty_\mu(X, \mathcal{A}, F)$.

The space $L^2_\mu(X, \mathcal{A}, F)$ is particularly interesting because if $F$ is a Hilbert space, then it can be given a Hilbert space structure which uses the inner product on $F$ (not quite, because the Hermitian form $\langle -, - \rangle_\mu$ that we obtain is not positive definite, which means that $\langle f, f \rangle_\mu = 0$ does necessarily imply that $f = 0$).

Let us start with the simple case where $F = \mathbb{C}$. If $f: X \to \mathbb{C}$ is a complex-valued function, then by $|f|^2$ we mean the function defined such that

\[ |f|^2(x) = f(x)\overline{f(x)} \text{ for all } x \in X. \]

For any two functions $f, g: X \to \mathbb{C}$, by $\langle f, g \rangle$ we mean the function defined such that

\[ \langle f, g \rangle(x) = f(x)\overline{g(x)} \text{ for all } x \in X. \]

For the more general case where $F$ is a Hilbert space with inner product $\langle -, - \rangle$, for any two functions $f, g: X \to F$, then by $\langle f, g \rangle$ we mean the function defined such that

\[ \langle f, g \rangle(x) = \langle f(x), g(x) \rangle \text{ for all } x \in X. \]

In particular,

\[ \|f\|^2 = \langle f, f \rangle. \]

From now on, when dealing with $L^2_\mu(X, \mathcal{A}, F)$, we assume that $F$ is a Hilbert space (over $\mathbb{C}$). If the reader feels more comfortable, he/she may assume that $F = \mathbb{C}$, but significant simplifications do not arise.

Proposition 4.40. The set $L^2_\mu(X, \mathcal{A}, F)$ is a vector space. For any two maps $f, g \in L^2_\mu(X, \mathcal{A}, F)$, we have $\langle f, g \rangle \in L^1_\mu(X, \mathcal{A}, \mathbb{C})$, and the map

\[ (f, g) \mapsto \int \langle f, g \rangle \, d\mu \]

is a Hermitian positive map (not necessarily positive definite).
4.7. THE SPACES $L^p_\mu(X, A, F)$ AND $L^p_\mu(X, A, F); p = 1, 2, \infty$

Proof. It is easy to see that $\langle f, g \rangle$ is a limit of step maps a.e., so $\langle f, g \rangle \in M_\mu(X, A, \mathbb{C})$, and we have the standard inequality

$$2|\langle f, g \rangle| \leq \|f\|^2 + \|g\|^2,$$

with $\|f\|^2 + \|g\|^2 \in L^1_\mu(X, A, \mathbb{R})$ since $f, g \in L^2_\mu(X, A, F)$. By Theorem 4.34, we have $\langle f, g \rangle \in L^1_\mu(X, A, \mathbb{C})$, so $\int \langle f, g \rangle d\mu$ is well defined. If $f \in L^2_\mu(X, A, F)$ and $g \in L^2_\mu(X, A, F)$, then since $\|f + g\|^2 \leq \|f\|^2 + 2|\langle f, g \rangle| + \|g\|^2$, as all the functions on the right-hand side are in $L^1_\mu(X, A, \mathbb{R})$, we have $f + g \in L^2_\mu(X, A, F)$. For any $\lambda \in \mathbb{C}$, we have $|\alpha|^2 \|f\| \in L^1_\mu(X, A, \mathbb{R})$, so $\lambda f \in L^2_\mu(X, A, F)$. Thus, $L^2_\mu(X, A, F)$ is a vector space. Using the linearity of the integral, it is easy to check that the map

$$(f, g) \mapsto \int \langle f, g \rangle d\mu$$

is a Hermitian positive map.

Definition 4.15. For any two functions $f, g \in L^2_\mu(X, A, F)$, the Hermitian map $\langle f, g \rangle_\mu$ is defined by

$$\langle f, g \rangle_\mu = \int \langle f, g \rangle d\mu$$

The $L^2$-semi-norm $\|f\|_2$ is given by

$$\|f\|_2 = \sqrt{\langle f, f \rangle_\mu} = \left( \int \langle f, f \rangle d\mu \right)^{1/2} = \left( \int \|f\|^2 d\mu \right)^{1/2}.$$

It is a standard result of linear algebra that the Cauchy–Schwarz inequality holds:

$$|\langle f, g \rangle_\mu| \leq \|f\|_2 \|g\|_2.$$

As a consequence $\| \cdot \|_2$ is a semi-norm.

Proposition 4.41. For any $f \in L^2_\mu(X, A, F)$, we have $\|f\|_2 = 0$ iff $f = 0$ a.e.

Proof. If $f = 0$ a.e., then $\langle f, f \rangle = 0$ a.e., so $\|f\|_2^2 = \int \langle f, f \rangle d\mu = 0$. Conversely, if $\|f\|_2 = 0$, then this means that $\int \langle f, f \rangle d\mu = 0$, but $\langle f, f \rangle \in L^1_\mu(X, A, \mathbb{R})$ is a positive function, so we know from Proposition 4.26 that $\langle f, f \rangle = 0$ a.e., that is, $f = 0$ a.e.

If $X$ has finite measure, then $L^2_\mu(X, A, F)$ is contained in $L^1_\mu(X, A, F)$.

Proposition 4.42. If $X$ has finite measure, then for any $f \in L^2_\mu(X, A, F)$, we have $\|f\|_1 \leq \|f\|_2 \|1_X\|_2$, and $L^2_\mu(X, A, F)$ is contained in $L^1_\mu(X, A, F)$.
CHAPTER 4. INTEGRATION

Proof. We apply the Cauchy–Schwarz inequality to \( \|f\| \) and to the constant function \( 1_X \) equal to 1 on \( X \). To be more specific, since

\[
\langle \|f\|, 1_X \rangle = \|f\| \int_X = \|f\|
\]

we have

\[
\langle \|f\|, 1_X \rangle_\mu = \int \langle \|f\|, 1_X \rangle d\mu = \int \|f\| = \|f\|_1,
\]

and

\[
\|\|f\||_2^2 = \langle \|f\|, \|f\| \rangle_\mu = \int \langle \|f\|, \|f\| \rangle d\mu = \int \|f\|^2 d\mu = \|f\|_2^2,
\]

and so the Cauchy–Schwarz inequality implies that \( \|f\|_1 \leq \|f\|_2 \|1_X\|_2 \). Obviously, this inequality shows that \( L^2_\mu(X, A, F) \) is contained in \( L^1_\mu(X, A, F) \).

It should be noted that if \( X \) has finite measure then the inclusion can be strict, and if \( X \) has infinite measure, then in general there are no inclusion properties.

Example 4.2.

1. If \( X = (0, 1) \), with the Lebesgue measure, then \( \frac{1}{\sqrt{x}} \in L^1((0, 1), \mu_L) \) but \( \frac{1}{\sqrt{x}} \notin L^2((0, 1), \mu_L) \).

2. If \( X = (1, \infty) \) with the Lebesgue measure, then \( \frac{1}{x} \in L^2((1, \infty), \mu_L) \) but \( \frac{1}{x} \notin L^1((1, \infty), \mu_L) \).

3. If \( X = (0, \infty) \) with the Lebesgue measure, then \( \frac{1}{(x+1)\sqrt{x}} \in L^1((0, \infty), \mu_L) \) but \( \frac{1}{(x+1)\sqrt{x}} \notin L^2((0, \infty), \mu_L) \).

One of the main properties of \( L^2_\mu(X, A, F) \) is that it is complete for the semi-norm \( \| \|_2 \). By taking the quotient of \( L^2_\mu(X, A, F) \) by the space of function equal to 0 a.e., we obtain a Hilbert space.

Theorem 4.43. Let \( (f_n)_{n\geq 1} \) be an \( \| \|_2 \)-Cauchy sequence of functions \( f_n \in L^2_\mu(X, A, F) \). Then there is a function \( f \in L^2_\mu(X, A, F) \) with the following properties:

1. The sequence \( (f_n)_{n\geq 1} \) converges to \( f \) in the \( \| \|_2 \)-semi-norm. Thus \( L^2_\mu(X, A, F) \) is complete.

2. There is a subsequence \( (f_{n_k})_{k\geq 1} \) of \( (f_n)_{n\geq 1} \) with the following properties:

   2. The subsequence \( (f_{n_k})_{k\geq 1} \) converges pointwise a.e. to \( f \).

3. For every \( \epsilon > 0 \), there is a subset \( Z \) such that \( \mu(Z) < \epsilon \) and the subsequence \( (f_{n_k})_{k\geq 1} \) converges uniformly to \( f \) on \( X - Z \).
4.7. THE SPACES $\mathcal{L}_p^\mu(X, \mathcal{A}, F)$ AND $\mathcal{L}_p^\mu(X, \mathcal{A}, F)$; $p = 1, 2, \infty$

A proof of Theorem 4.43 is given in Lang [24] (Chapter VII, §1).

In view of Proposition 4.41, we make the following definition.

**Definition 4.16.** Let $L^2_\mu(X, \mathcal{A}, F)$ be the quotient of the vector space $L^2_\mu(X, \mathcal{A}, F)$ by the subspace of functions equal to 0 a.e. (which is the set of functions $f$ such that $\|f\|_2 = 0$). The norm induced by the semi-norm $\|\cdot\|_2$ on $L^2_\mu(X, \mathcal{A}, F)$ is called the $L^2$-norm.

Obviously the positive Hermitian form $\langle f, g \rangle_\mu$ induces a positive definite Hermitian form on $L^2_\mu(X, \mathcal{A}, F)$. Theorem 4.43 immediately implies the following result.

**Theorem 4.44.** (Fischer–Riesz) The space $L^2_\mu(X, \mathcal{A}, F)$ is a Hilbert space under the positive definite Hermitian form induced by $\langle -,- \rangle_\mu$.

**Definition 4.17.** The norm $\|\cdot\|_2$ associated with the inner product $\langle -,- \rangle_\mu$ on $L^2_\mu(X, \mathcal{A}, F)$ is called the $L^2$-norm.

We will show shortly that the space of $\mu$-step functions is dense in $L^2_\mu(X, \mathcal{A}, F)$ (for the $L^2$-norm). First here is a corollary of Theorem 4.44.

**Proposition 4.45.** If $(f_n)_{n \geq 1}$ is a $\|\cdot\|_2$-Cauchy sequence of functions $f_n \in L^2_\mu(X, \mathcal{A}, F)$, and if $(f_n)_{n \geq 1}$ converges pointwise a.e. to a function $f : X \to F$, then $f \in L^2_\mu(X, \mathcal{A}, F)$, and $(f_n)_{n \geq 1}$ converges to $f$ in the $\|\cdot\|_2$-semi-norm.

The Lebesgue dominated convergence theorem also holds for $L^2_\mu(X, \mathcal{A}, F)$.

**Theorem 4.46.** (Lebesgue Dominated Convergence Theorem for $L^2_\mu$) Let $(f_n)_{n \geq 1}$ be a sequence of functions $f_n \in L^2_\mu(X, \mathcal{A}, F)$. If $(f_n)$ converges pointwise a.e. to a function $f : X \to F$, and if there is some function $g \in L^1(X, \mathcal{A}, \mathbb{R})$ such that $g \geq 0$ and $\|f_n\| \leq g$ for all $n \geq 1$, then $f \in L^2_\mu(X, \mathcal{A}, F)$ and $(f_n)_{n \geq 1}$ converges to $f$ in the $\|\cdot\|_2$-norm. Consequently

$$\lim_{n \to \infty} \int f_n d\mu = \int f d\mu.$$  

A proof of Theorem 4.46 is given in Lang [24] (Chapter VII, §1).

Theorem 4.34 also holds for $L^2_\mu(X, \mathcal{A}, F)$.

As a corollary of Theorem 4.46 we can show that the $\mu$-step functions are dense in $L^2_\mu(X, \mathcal{A}, F)$.

**Proposition 4.47.** The subspace $\text{Step}_\mu(X, \mathcal{A}, F)$ is dense in $L^2_\mu(X, \mathcal{A}, F)$ with respect to the $L^2$-semi-norm.
Proof. Let \( f \in L^2_\mu(X, \mathcal{A}, F) \). Since \( f \) is \( \mu \)-measurable, there is a sequence \((f_n)_{n \geq 1}\) of \( \mu \)-step functions \( f_n \) that converges pointwise a.e. to \( f \). For every \( n \geq 1 \) and every \( x \in X \), define \( g_n \) by

\[
g_n(x) = \begin{cases} 
  f_n(x) & \text{if } \|f_n(x)\| \leq 2 \|f(x)\| \\
  0 & \text{if } \|f_n(x)\| > 2 \|f(x)\|.
\end{cases}
\]

We may assume that \( f \) is measurable since it differs from a measurable function on a set of measure zero. Then the functions \( g_n \) are \( \mu \)-step functions, they satisfy the inequality \( \|g_n\| \leq 2 \|f\| \) with \( 2 \|f\| \in L^2_\mu(X, \mathcal{A}, \mathbb{R}) \), and the sequence \((g_n)\) converges a.e. to \( f \). By Theorem 4.46, the sequence \((g_n)\) converges to \( f \) in the \( \|\|_2 \)-norm, which proves that \( \text{Step}_\mu(X, \mathcal{A}, F) \) is dense in \( L^2_\mu(X, \mathcal{A}, F) \) with respect to the \( L^2 \)-semi-norm.

We now would like to understand the duals of \( L^1_\mu(X, \mathcal{A}, F) \) and \( L^2_\mu(X, \mathcal{A}, F) \), that is, the spaces of continuous linear forms on \( L^1_\mu(X, \mathcal{A}, F) \) and \( L^2_\mu(X, \mathcal{A}, F) \) (with values in \( \mathbb{C} \)). In the case of \( L^1_\mu(X, \mathcal{A}, F) \), it is a classical theorem (the Riesz representation theorem) that the dual of a Hilbert space is isomorphic to itself, so the dual of \( L^2_\mu(X, \mathcal{A}, F) \) is isomorphic to \( L^2_\mu(X, \mathcal{A}, F) \). In the case of \( L^1_\mu(X, \mathcal{A}, F) \), it turns out that its dual is isomorphic to a space denoted \( L^\infty_\mu(X, \mathcal{A}, F) \). Here we assumed that \( F \) is a Hilbert space.

To define \( L^\infty_\mu(X, \mathcal{A}, F) \), we only need the fact that \( F \) is a Banach space. The space \( L^\infty_\mu(X, \mathcal{A}, F) \) consists of all functions \( f : X \to F \) that are equal to a bounded \( \mu \)-measurable function a.e. We can define a semi-norm on \( L^\infty_\mu(X, \mathcal{A}, F) \) as follows.

**Definition 4.18.** For any function \( f \in M_\mu(X, \mathcal{A}, F) \), define the essential sup or semi-norm \( \|f\|_\infty \) of \( f \) by

\[
\|f\|_\infty = \inf \{ \alpha \in \mathbb{R}_+ : \mu(\{x \in X \mid \|f(x)\| \geq \alpha\}) = 0 \}.
\]

The space \( L^\infty_\mu(X, \mathcal{A}, F) \) is the set of functions \( f \in M_\mu(X, \mathcal{A}, F) \) such that \( \|f\|_\infty < +\infty \).

The definition of \( \|f\|_\infty \) makes is clear that \( \|f\|_\infty = 0 \) if \( f = 0 \) a.e. Observe that \( \|f\|_\infty \) is the greatest lower bound of the numbers \( \alpha \geq 0 \) such that a \( \mu \)-measurable function \( f \) has the property that \( \|f(x)\| \geq \alpha \) on a set of measure zero, in other words, such a \( \mu \)-measurable function is bounded a.e.

The space \( L^\infty_\mu(X, \mathcal{A}, F) \) is a vector space. We also have the following result showing that \( L^\infty_\mu(X, \mathcal{A}, F) \) is complete in the semi-norm \( \|\|_\infty \), but unless \( X \) has finite measure, the \( \mu \)-step map are not dense in \( L^\infty_\mu(X, \mathcal{A}, F) \).

**Theorem 4.48.** The following properties hold.

1. The space \( L^\infty_\mu(X, \mathcal{A}, F) \) is complete in the semi-norm \( \|\|_\infty \). Furthermore, if \((f_n)_{n \geq 1}\) is an \( \|\|_\infty \)-Cauchy sequence, then there is a set \( Z \) of measure zero such that \((f_n)_{n \geq 1}\) converges uniformly to \( f \) on \( X - Z \).

2. If \( F \) is finite-dimensional, then the step maps (not the \( \mu \)-step maps) are dense in \( L^\infty_\mu(X, \mathcal{A}, F) \).
3. If $X$ has finite measure, then for every $\epsilon > 0$ and every $f \in \mathcal{L}_p^\infty(X, \mathcal{A}, F)$, there is a $\mu$-step map $s$ and a subset $Z$ with $\mu(Z) < \epsilon$ such that

$$\|f - s\| < \epsilon \quad \text{on} \quad X - Z.$$ 

Note that the constant with value 1 belongs to $\mathcal{L}_\infty^\infty(X, \mathcal{A}, F)$, so if $X$ has infinite measure, there is no way that it is a uniform limit of $\mu$-step maps, since a $\mu$-step map vanishes outside of a set of finite measure.

**Remark:** If $X$ has finite measure, then we have the inclusion $\mathcal{L}_\infty^\infty(X, \mathcal{A}, F) \subseteq \mathcal{L}_p^2(X, \mathcal{A}, F)$. In fact, $\mathcal{L}_\infty^\infty(X, \mathcal{A}, F) \subseteq \mathcal{L}_p^q(X, \mathcal{A}, F)$ for all $p, q \geq 1$ with $p > q$; see Marle [27] (Chapter 4, Proposition 4.5.7).

**Definition 4.19.** Let $\mathcal{L}_\infty^\infty(X, \mathcal{A}, F)$ be the quotient of the vector space $\mathcal{L}_\infty^\infty(X, \mathcal{A}, F)$ by the subspace of functions equal to 0 a.e. (which is the set of functions $f$ such that $\|f\|_\infty = 0$). The norm induced by the semi-norm $\| \|_\infty$ on $\mathcal{L}_\infty^\infty(X, \mathcal{A}, F)$ is called the $L^\infty$-norm.

It should be noted that both the monotone convergence theorem and the dominated convergence theorem fail for $\mathcal{L}_\infty^\infty(X, \mathcal{A}, F)$. Convergence in the $\| \|_\infty$-semi-norm fails; see Marle [27] (Chapter 4, Section 2).

We now consider the duality between the spaces $\mathcal{L}_1^1(X, \mathcal{A}, F)$ and $\mathcal{L}_\infty^\infty(X, \mathcal{A}, F)$. The field $\mathbb{C}$ is a Hilbert space, but for the general case we need to assume that $F$ is a Hilbert space. The key point is that by Proposition 4.35(2), for any $f \in \mathcal{L}_1^1(X, \mathcal{A}, F)$ and any $g \in \mathcal{L}_\infty^\infty(X, \mathcal{A}, F)$, then $(f, g) \in \mathcal{L}_1^1(X, \mathcal{A}, F)$.

**Definition 4.20.** For any functions $f \in \mathcal{L}_1^1(X, \mathcal{A}, F)$ and $g \in \mathcal{L}_\infty^\infty(X, \mathcal{A}, F)$, define $[f, g]_\mu$ by

$$[f, g]_\mu = \int (f, g) d\mu.$$ 

We obtain a map

$$[-, -]_\mu : \mathcal{L}_1^1(X, \mathcal{A}, F) \times \mathcal{L}_\infty^\infty(X, \mathcal{A}, F) \to \mathbb{C}$$

which is a sesquilinear pairing.

For simplicity, let us consider the special case where $F = \mathbb{C}$. In this case, we can define a bilinear (as opposed to sesquilinear) pairing $[-, -]_\mu : \mathcal{L}_1^1(X, \mathcal{A}, \mathbb{C}) \times \mathcal{L}_\infty^\infty(X, \mathcal{A}, \mathbb{C}) \to \mathbb{C}$ given by

$$[f, g]_\mu = \int fg d\mu.$$ 

Observe that we intentionally used $fg$ instead of $f\bar{g}$, because we simply want a bilinear pairing.
Whenever we have a bilinear pairing $\varphi : E \times F \to \mathbb{C}$, recall that we define the linear maps $l_\varphi : E \to F^*$ and $r_\varphi : F \to E^*$ such that, for every $u \in E$,

$$l_\varphi(u)(y) = \varphi(u, y) \quad \text{for all } y \in F,$$

and for every $v \in F$,

$$r_\varphi(v)(x) = \varphi(x, v) \quad \text{for all } x \in E.$$

**Definition 4.21.** A bilinear pairing $\varphi$ is nondegenerate if for every $u \in E$, if $\varphi(u, v) = 0$ for all $v \in F$, then $u = 0$, and for every $v \in F$, if $\varphi(u, v) = 0$ for all $u \in E$, then $v = 0$.

Then if $\varphi$ is nondegenerate, then the maps $l_\varphi$ and $r_\varphi$ are injective. They are not surjective in general.

If $E$ is a normed vector space, then its dual $E'$ is the space of all continuous linear maps from $E$ to $\mathbb{C}$. We have $E' \subseteq E^*$, and the inclusion is strict if $E$ is infinite-dimensional.

The following result holds. For simplicity of notation, we drop $\varphi$ when writing $l_\varphi$ and $r_\varphi$.

**Theorem 4.49.** Assume $(X, \mathcal{A}, \mu)$ is a measure space and that $\mu$ is $\sigma$-finite. Then the bilinear pairing $[-, -]_\mu : L^1_\mu(X, \mathcal{A}, \mathbb{C}) \times L^\infty_\mu(X, \mathcal{A}, \mathbb{C}) \to \mathbb{C}$ is nondegenerate. It satisfies the inequality

$$|[f, g]_\mu| \leq \|fg\|_1 \leq \|f\|_1 \|g\|_\infty.$$

The map $l$ is a norm-preserving injective linear map between $L^1_\mu(X, \mathcal{A}, \mathbb{C})$ and the dual $L^\infty_\mu(X, \mathcal{A}, \mathbb{C})'$ of $L^\infty(X, \mathcal{A}, \mathbb{C})$, and the map $r$ is a norm-preserving injective linear between $L^\infty_\mu(X, \mathcal{A}, \mathbb{C})$ and the dual $L^1_\mu(X, \mathcal{A}, \mathbb{C})'$ of $L^1_\mu(X, \mathcal{A}, \mathbb{C})$. Furthermore, the map $r : L^\infty_\mu(X, \mathcal{A}, \mathbb{C}) \to L^1_\mu(X, \mathcal{A}, \mathbb{C})'$ is an isomorphism.

A proof of Theorem 4.49 is given in Lang [24] (Chapter VII, §2). Theorem 4.49 can be generalized to a Hilbert space $F$, one just has to exercise caution in defining $l$ and $r$ to deal with sequilinearity.

The map $l : L^1_\mu(X, \mathcal{A}, \mathbb{C}) \to L^\infty_\mu(X, \mathcal{A}, \mathbb{C})'$ is not surjective, and understanding which linear forms in $L^\infty_\mu(X, \mathcal{A}, \mathbb{C})'$ can be represented by functions in $L^1_\mu(X, \mathcal{A}, \mathbb{C})$ is a natural question. A partial answer to this question is the Radon–Nikodym theorem, but will this would lead us too far. The interested reader is referred to Lang [24] or Rudin [36].

### 4.8 Products of Measure Spaces and Fubini’s Theorem

The purpose of this section is to define the notion of product measure space and product measure, given two measure spaces $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$. Then we will state Fubini’s theorem (also known as the theorem of Lebesgue–Fubini), which allows us to compute the...
We begin by recalling what we did in Example 3.1. We defined the set $\mathcal{R}$ of rectangles in $X \times Y$ as follows:

$$\mathcal{R} = \{A \times B \in X \times Y \mid A \in \mathcal{A}, B \in \mathcal{B}\}.$$ 

The set $\mathcal{R}$ is a semi-algebra, and it can be shown that the set $\mathcal{B}(\mathcal{R})$ of finite unions of pairwise disjoint sets in $\mathcal{R}$ is the smallest algebra containing the semi-algebra $\mathcal{R}$.

**Definition 4.22.** Let $\mathcal{A} \otimes \mathcal{B}$ be the smallest $\sigma$-algebra generated by $\mathcal{R}$ (and thus by $\mathcal{B}(\mathcal{R})$); see Proposition 3.3.4

The hard part is now to define a product measure $\lambda$ on $\mathcal{A} \otimes \mathcal{B}$ which satisfies the natural identity

$$\lambda(A \times B) = \mu(A)\nu(B)$$

for all rectangles $A \times B$. Here as in Section 3.1 we use extended multiplication on $\mathbb{R}_+$, where

$$a \cdot (+\infty) = (+\infty) \cdot a = +\infty$$

if $0 < a \leq +\infty$, and

$$0 \cdot (+\infty) = (+\infty) \cdot 0 = 0.$$

We need a few definitions.

**Definition 4.23.** Given any subset $E \subseteq X \times Y$, for any $x \in X$, we define the section of $E$ (determined by $x$) as the subset $E_x$ given by

$$E_x = \{y \in Y \mid (x,y) \in E\} \subseteq Y.$$ 

Similarly, for any $y \in Y$, we define the section of $E$ (determined by $y$) as the subset $E_y$ given by

$$E_y = \{x \in X \mid (x,y) \in E\} \subseteq X.$$ 

See Figure 4.9.

**Proposition 4.50.** The sections of any subset $E \in \mathcal{A} \otimes \mathcal{B}$ are measurable.

**Proof idea.** Let $\mathcal{E}$ be the family of subsets of $X \times Y$ defined as follows:

$$\mathcal{E} = \{F \subseteq X \times Y \mid F_x \in \mathcal{B} \text{ for all } x \in X, \text{ and } F_y \in \mathcal{A} \text{ for all } y \in Y\}.$$ 

These are the subsets of $X \times Y$ whose sections are measurable. Then prove that $\mathcal{E}$ is a $\sigma$-algebra containing $\mathcal{R}$, which implies that $\mathcal{E} = \mathcal{A} \otimes \mathcal{B}$. 

---

4 The meaning of the tensor sign $\otimes$ in the notation $\mathcal{A} \otimes \mathcal{B}$ is a completely different from its meaning in a tensor product of vector spaces. Hopefully, the two notions will never appear together!
Figure 4.9: Let $X = Y = \mathbb{R}$. The top figure illustrates an $x$-section of the peach set $E$, while the bottom figure illustrates a $y$-section.

**Definition 4.24.** Given any function $f : X \times Y \to \mathbb{R}$, for any $x \in X$, we define the *section of $f$ (determined by $x$)* as the function $f_x : Y \to \mathbb{R}$ given by

$$f_x(y) = f(x, y) \quad \text{for all } y \in Y.$$  

Similarly, for any $y \in Y$, we define the *section of $f$ (determined by $y$)* as the function $f_y : X \to \mathbb{R}$ given by

$$f_y(x) = f(x, y) \quad \text{for all } x \in X.$$  

**Proposition 4.51.** If $f : X \times Y \to \mathbb{R}$ is a measurable function (on $(X \times Y, \mathcal{A} \otimes \mathcal{B})$, then every section of $f$ is measurable.

**Proof.** By Proposition 3.12, it suffices to show that the inverse image of every open subset of the form $(\alpha, \infty)$ is measurable.

For any $x \in X$, for any $\alpha \in \mathbb{R}$, we have

$$\{y \in Y \mid f_x(y) < \alpha\} = \{y \in Y \mid f(x, y) < \alpha\} = \{(x, y) \in X \times Y \mid f(x, y) < \alpha\}_x,$$

and this last subset is measurable by Proposition 4.50. The proof for $f_y$ is similar. \qed
Definition 4.25. Given an algebra \( A \) of sets, a measure on \( A \) satisfies the same axioms as a measure on a \( \sigma \)-algebra; see Definition 3.9.

The next two results take a lot more work.

Proposition 4.52. Let \((X, \mathcal{A}, \mu)\) and \((Y, \mathcal{B}, \nu)\) be two measure spaces, and assume that \(\mu\) and \(\nu\) are \(\sigma\)-finite. Then the map \(\lambda : \mathcal{R} \to [0, +\infty)\) given by
\[
\lambda(A \times B) = \mu(A)\nu(B)
\]
has a unique extension to a \(\sigma\)-finite measure on the algebra \(\mathcal{B}(\mathcal{R})\).

A proof of Proposition 4.52 can be found in course notes given by Philippe G. Ciarlet in 1970-1971 at ENPC (Paris, France). Interestingly, the proof uses the Monotone Convergence Theorem. A related treatment is given in Halmos [22] (Chapter VII); see also Lang [24] (Chapter VI, §8) and Marle [27] (Chapter 5, Section 2).

Theorem 4.53. Let \((X, \mathcal{A}, \mu)\) and \((Y, \mathcal{B}, \nu)\) be two measure spaces, and assume that \(\mu\) and \(\nu\) are \(\sigma\)-finite. Then the map \(\lambda : \mathcal{R} \to [0, +\infty)\) given by
\[
\lambda(A \times B) = \mu(A)\nu(B)
\]
has a unique extension to a measure \(\lambda = \mu \otimes \nu\) on the \(\sigma\)-algebra \(\mathcal{A} \otimes \mathcal{B}\). The measure \(\mu \otimes \nu\) is \(\sigma\)-finite.

The following properties hold for any measurable subset \(E \in \mathcal{A} \times \mathcal{B}\):

1. We have
\[
(\mu \otimes \nu)(E) = \inf \left\{ \sum_{i=1}^{\infty} \mu(A_i)\nu(B_j) \mid E \subseteq \bigcup_{i=1}^{\infty} (A_i \times B_i), A_i \in \mathcal{A}, B_j \in \mathcal{B} \right\}. \tag{*}
\]

2. The map \(\nu_E\) from \(X\) to \(\mathbb{R}_+\) given by \(x \mapsto \nu(E_x)\) is measurable (w.r.t. \(\mathcal{A}\)), and the map \(\mu_E\) from \(Y\) to \(\mathbb{R}_+\) given by \(y \mapsto \mu(E_y)\) is measurable (w.r.t. \(\mathcal{B}\)). One of these maps is integrable iff the other is integrable.

3. We have
\[
(\mu \otimes \nu)(E) = \begin{cases} 
\int \nu_E \, d\mu = \int \mu_E \, d\nu & \text{if both } \nu_E \text{ and } \mu_E \text{ are integrable} \\
+\infty & \text{otherwise}.
\end{cases} \tag{**}
\]

A proof of Theorem 4.53 can be found in course notes given by Philippe G. Ciarlet in 1970-1971 at ENPC (Paris, France). Again, the proof uses the Monotone Convergence
Theorem. A related treatment is given in Halmos [22] (Chapter VII); see also Lang [24] (Chapter VI, §8) and Marle [27] (Chapter 5, Section 2, Proposition 5.2.3).

If \((X, \mathcal{A}, \mu)\) and \((Y, \mathcal{B}, \nu)\) are two measure spaces with \(\mu\) and \(\nu\) both \(\sigma\)-finite. Then for any Banach space \(F\), we have the space of integrable functions \(L_{\mu \otimes \nu}(X \times Y, \mathcal{A} \otimes \mathcal{B}, F)\). The problem is to find a way to compute an integral \(\iint f \, d(\mu \otimes \nu)\), also written \(\iint f \, d\mu \otimes d\nu\), as two successive integrals. The answer is given by a theorem known as Fubini’s theorem. The first version of this theorem was proved by Lebesgue, and then was generalized by Fubini. For this reason some authors refer to this theorem as the Lebesgue–Fubini theorem, but it seems more common to call it simply Fubini’s theorem.

**Theorem 4.54.** (Fubini’s Theorem, Part 1) Let \((X, \mathcal{A}, \mu)\) and \((Y, \mathcal{B}, \nu)\) be two measure spaces with \(\mu\) and \(\nu\) both \(\sigma\)-finite. Consider a function \(f : X \times Y \to F\), where \(F\) is a Banach space. If \(f \in L_{\mu \otimes \nu}(X \times Y, \mathcal{A} \otimes \mathcal{B}, F)\) then:

1. the section \(f_x : Y \to F\) is \(\nu\)-integrable for almost all \(x \in X\), the section \(f_y : X \to F\) is \(\mu\)-integrable for almost all \(y \in Y\).

2. The map from \(X\) to \(F\) defined a.e. by

\[
x \mapsto \int f_x \, d\nu
\]

is \(\mu\)-integrable, and the map from \(Y\) to \(F\) defined a.e. by

\[
y \mapsto \int f_y \, d\mu
\]

is \(\nu\)-integrable.

Then

\[
\iint f \, d\mu \otimes d\nu = \int_X \left( \int_Y f_x \, d\nu \right) \, d\mu = \int_Y \left( \int_X f_y \, d\mu \right) \, d\nu.
\]

Theorem 4.54 is proved in Marle [27] (Chapter 5, Section 2, Theorem 5.2.10), and Lang [24] (Chapter VI, §8, Theorem 8.4).

Theorem 4.54 assumes that \(f\) is integrable. It is possible to weaken this assumption at the price of strengthening the other conditions. However, this is worth it in practice.

**Theorem 4.55.** (Fubini’s Theorem, Part 2) Let \((X, \mathcal{A}, \mu)\) and \((Y, \mathcal{B}, \nu)\) be two measure spaces with \(\mu\) and \(\nu\) both \(\sigma\)-finite. Consider a function \(f : X \times Y \to F\), where \(F\) is a Banach space. If \(f \in \mathcal{M}_{\mu \otimes \nu}(X \times Y, \mathcal{A} \otimes \mathcal{B}, F)\) and if the following conditions hold:

1. the section \(f_x : Y \to F\) is \(\nu\)-integrable for almost all \(x \in X\), the section \(f_y : X \to F\) is \(\mu\)-integrable for almost all \(y \in Y\).
2. The map from $X$ to $\mathbb{R}$ defined a.e. by

$$x \mapsto \int \|f_x\| \, d\nu$$

is $\mu$-integrable, and the map from $Y$ to $\mathbb{R}$ defined a.e. by

$$y \mapsto \int \|f_y\| \, d\mu$$

is $\nu$-integrable.

Then $f \in L_{\mu \otimes \nu}(X \times Y, A \otimes B, F)$ and

$$\int \int f \, d\mu \otimes d\nu = \int_X \left( \int_Y f_x \, d\nu \right) \, d\mu = \int_Y \left( \int_X f_y \, d\mu \right) \, d\nu.$$

Theorem 4.55 is proved in Lang [24] (Chapter VI, §8, Theorem 8.7); see also Marle [27] (Chapter 5, Section 2).

In practice, it is customary to use a less formal notation to express Fubini’s theorem, namely

$$\int \int f(x, y) \, d\mu(x) \otimes d\nu(y) = \int_X \left( \int_Y f(x, y) \, d\nu(y) \right) \, d\mu(x) = \int_Y \left( \int_X f(x, y) \, d\mu(x) \right) \, d\nu(y),$$

and the measure $d\mu(x) \otimes d\nu(y)$ is often denoted simply by $d\mu(x) d\nu(y)$. 
Chapter 5

Radon Functionals and Radon Measures on Locally Compact Spaces

After having considered a very general theory of integration of functions defined on an arbitrary measure space and taking their values in any Banach space, we turn to the special case of complex-valued or real-valued functions defined on a locally compact space \( X \). This corresponds to measure spaces \((X, \mathcal{B}, \mu)\), where \( X \) is a locally compact space, \( \mathcal{B} \) is the \( \sigma \)-algebra of Borel set (which is the smallest \( \sigma \)-algebra containing the open subsets of \( X \)), and \( \mu \) is any (positive) measure on \( \mathcal{B} \), which we call a Borel measure.

The theme of this chapter is that a Borel measure \( \mu \) can be used to define linear forms on various function spaces. For example, pick the space \( K_C(X) \) of continuous functions on \( X \) with compact support. For every function \( f \in K_C(X) \) we can compute the integral

\[
\varphi_\mu(f) = \int f \, d\mu.
\]

We have to check that functions in \( f \in K_C(X) \) are integrable, which is indeed true. We obtain a map \( \varphi_\mu : K_C(X) \to \mathbb{C} \), and since the integral is a linear operator, the map \( \varphi_\mu \) is linear. In general it is not continuous, but it satisfies some weaker continuity properties. It is also a positive map, which means that \( \Phi(f) \geq 0 \) for every positive function \( f \geq 0 \).

What F. Riesz and J. Radon and discovered is that, in some sense to be made precise, a special class of Borel measures are in one-to-one correspondence with the positive linear forms on the space \( K_C(X) \). This means that for every positive linear form \( \Phi \) on \( K_C(X) \), there is a (unique) Borel measure \( m_\Phi \) with some special properties such that \( \Phi \) is represented by \( m_\Phi \), in the sense that

\[
\Phi(f) = \int f \, dm_\Phi \quad \text{for all } f \in K_C(X).
\]

There are two versions of this correspondence theorem known as the Radon–Riesz theorem, depending on the conditions imposed on the Borel measures.
These results are similar in flavor to the fact known from linear algebra that, in a finite-dimensional vector space $E$ with an inner product $\langle -,- \rangle$, every linear form $\varphi \in E^*$ is represented by a unique vector $u \in E$, in the sense that
\[ \varphi(v) = \langle v, u \rangle \quad \text{for all } v \in E. \]

If $(E, \langle -,- \rangle)$ is an infinite-dimensional vector space which is a Hilbert space (it is complete for the norm $u \mapsto \sqrt{\langle u,u \rangle}$), then by the Riesz representation theorem, every continuous linear form $\varphi \in E'$ is represented by a unique vector $u \in E$, in the sense that
\[ \varphi(v) = \langle v, u \rangle \quad \text{for all } v \in E. \]

The Radon–Riesz theorems show that certain kinds of (possibly discontinuous) linear forms on $K_C(X)$ can be represented using integration instead of an inner product.

The main limitation of this approach is that the linear forms $\Phi$ induced by a positive measure are positive, which means that $\Phi(f) \geq 0$ if $f \geq 0$. In particular, it impossible to represent an arbitrary continuous linear form on $K_C(X)$ using integration. The solution to overcome this limitation is to generalize the notion of measure so that a measure can take negative, or even complex, values! We will show how to do this. We will also see that, in the end, complex measures can be expressed in terms of four positive measures, but these positive measures only take finite values in $\mathbb{R}_+$. Then we will obtain a third Radon–Riesz correspondences between the continuous linear forms on $K_C(X)$ and certain kinds of complex Borel measures. This correspondence plays a crucial role in defining the notion of convolution on a locally compact group.

In this chapter every topological space $X$ is assumed to be locally compact (and Hausdorff).

## 5.1 Positive Radon Functionals Induced by Borel Measures

One direction of the correspondence (Borel measures $\Rightarrow$ linear forms) is easy to describe. It is the observation that the linear forms induced by Borel measures are positive.

**Definition 5.1.** For any function $f : X \to \mathbb{C}$, we write $f \geq 0$ if $f(X) \subseteq [0, \infty)$. If $f, g : X \to \mathbb{R}$, we write $f \leq g$ iff $g - f \geq 0$. A linear form $\Phi : K_C(X) \to \mathbb{C}$ is positive if for every $f \in K_C(X)$, if $f \geq 0$, then $\Phi(f) \in \mathbb{R}$ and $\Phi(f) \geq 0$.

A positive linear form has the following properties.

**Proposition 5.1.** If $\Phi : K_C(X) \to \mathbb{C}$ is a positive linear form, then the following properties hold:

1. For any real-valued function $f \in K_R(X)$ we must have $\Phi(f) \in \mathbb{R}$.  

(2) For any two real-valued functions \( f, g \in K_\mathbb{R}(X) \), if \( f \leq g \), then \( \Phi(f) \leq \Phi(g) \).

Proof. Indeed, a real-valued function \( f \) can be written uniquely as \( f = f^+ - f^- \), with \( f^+, f^- \in K_\mathbb{R}(X) \), \( f^+ \geq 0 \) and \( f^- \geq 0 \). Since \( \Phi \) is linear,

\[
\Phi(f) = \Phi(f^+) - \Phi(f^-) \in \mathbb{R},
\]

since \( \Phi(f^+) \geq 0 \) and \( \Phi(f^-) \geq 0 \) as \( \Phi \) is positive.

We have \( f \leq g \) iff \( g - f \geq 0 \), and since \( \Phi \) is positive, \( \Phi(g - f) \geq 0 \), but since \( \Phi \) is linear and positive, \( \Phi(g) - \Phi(f) \geq 0 \) with \( \Phi(f), \Phi(g) \in \mathbb{R} \), that is, \( \Phi(f) \leq \Phi(g) \).

The following proposition yields the mapping from Borel measures to positive linear forms.

**Proposition 5.2.** Assume that the Borel measure \( \mu \) has the property that \( \mu(K) \) is finite for every compact subset of \( X \) (since \( X \) is Hausdorff, a compact set is closed, and thus a Borel set). Every function \( f \in K_\mathbb{C}(X) \) is integrable. Furthermore, the map \( \varphi_\mu : K_\mathbb{C}(X) \to \mathbb{C} \) given by

\[
\varphi_\mu(f) = \int f \, d\mu
\]

is a positive linear form.

Proof. Since \( f \) has compact support, say \( K \), and since it is continuous, it is bounded, say \( |f| \leq M \chi_K \). Since \( f \) is continuous, it is measurable, and the function \( M \chi_K \) is a step function which is integrable since \( \mu(K) \) is finite. By Theorem 4.34, the function \( f \) is integrable. By Proposition 4.23, the map \( \varphi_\mu \) is linear and positive.

Remark: As a point of terminology, the map \( \varphi_\mu : K_\mathbb{C}(X) \to \mathbb{C} \) just is just a linear form, but since its domain is a function space (\( K_\mathbb{C}(X) \)), it is customary to call it a linear functional.

The remarkable fact is that any positive linear functional \( \Phi : K_\mathbb{C}(X) \to \mathbb{C} \) determines a Borel measure \( m_\Phi \) (with some special properties) such that

\[
\Phi(f) = \int f \, dm_\Phi \quad \text{for all } f \in K_\mathbb{C}(X).
\]

Knowing how to integrate functions in \( K_\mathbb{C}(X) \) is sufficient to determine the measure \( m_\Phi \) completely. In some sense, continuous functions with compact support play the role of \( \mu \)-step functions.

Recall that for any compact subset \( K \) of \( X \), we denote by \( K(K; \mathbb{C}) \) the set of complex-valued continuous functions whose support is contained in \( K \) (and similarly \( K(K; \mathbb{R}) \) for real-valued functions). Interestingly, every positive linear functional on \( K_\mathbb{C}(X) \) is continuous on \( K(K; \mathbb{C}) \) for every compact subset \( K \) of \( X \).
Proposition 5.3. If \( \Phi: \mathcal{K}_C(X) \to \mathbb{C} \) is a positive linear functional on \( \mathcal{K}_C(X) \) then for every compact subset \( K \) of \( X \), there is some real number \( c_K \geq 0 \) such that \( |\Phi(f)| \leq c_K \|f\|_\infty \) for all \( f \in \mathcal{K}(K; \mathbb{C}) \).

Proof. Every function \( f \) in \( \mathcal{K}(K; \mathbb{C}) \) can be written uniquely as \( f = f_1 + if_2 \) with \( f_1, f_2 \in \mathcal{K}_\mathbb{R}(X) \). Since \( \Phi \) is positive linear, we have \( \Phi(f_1) \in \mathbb{R}, \Phi(f_2) \in \mathbb{R} \) and \( \Phi(f) = \Phi(f_1 + if_2) = \Phi(f_1) + i\Phi(f_2) \), so

\[
|\Phi(f)| = \sqrt{\Phi(f_1)^2 + \Phi(f_2)^2}.
\]

If we can show that \( |\Phi(f_1)| \leq c_1 \|f_1\|_\infty \) and \( |\Phi(f_2)| \leq c_2 \|f_2\|_\infty \), then we get

\[
|\Phi(f)| \leq \sqrt{c_1^2 \|f_1\|_\infty^2 + c_2^2 \|f_2\|_\infty^2} \leq \sqrt{c_1^2 + c_2^2} \|f\|_\infty,
\]

since

\[
\|f\|_\infty = \sup_{x \in K} |f(x)| = \sup_{x \in K} \sqrt{(f_1(x))^2 + (f_2(x))^2} = \sqrt{\sup_{x \in K} (f_1(x))^2 + \sup_{x \in K} (f_2(x))^2},
\]

which implies that

\[
\|f_1\|_\infty = \sup_{x \in K} |f_1(x)| \leq \|f\|_\infty,
\]

and

\[
\|f_2\|_\infty = \sup_{x \in K} |f_2(x)| \leq \|f\|_\infty.
\]

Therefore we may assume that \( f \in \mathcal{K}(K; \mathbb{R}) \). By Proposition 13.39, there is a continuous function with compact support \( g \in \mathcal{K}(K; \mathbb{R}) \) (a bump function) such that \( g(x) = 1 \) for all \( x \in K \). For any \( f \in \mathcal{K}_\mathbb{R}(X) \), we have

\[
-g \|f\|_\infty \leq f \leq g \|f\|_\infty
\]

and since \( \Phi \) is a positive linear functional, by Proposition 5.1(2), we get

\[
-\Phi(g) \|f\|_\infty \leq \Phi(f) \leq \Phi(g) \|f\|_\infty
\]

that is

\[
|\Phi(f)| \leq \Phi(g) \|f\|_\infty,
\]

as desired \( \square \)

Proposition 5.3 suggests that the linear functionals \( \Phi: \mathcal{K}_C(X) \to \mathbb{C} \) satisfying the conclusion of the proposition are of particular interest, and they are. In fact the measure theory and the integration theory for complex-valued functions on a locally compact space can be developed entirely in terms of these functionals. This approach is presented in Dieudonné [12], Bourbaki [3, 5, 6], and Schwartz [41]. Dieudonné and Bourbaki even go as far as calling such functionals measures, which we feel is unfortunate because this term already has a well
est. meaning. Unlike these two previous sources, Schwartz actually develops in parallel both the theory of integration using measure theory, and the theory of integration using certain linear functionals that he calls Radon measures. Again, we find this terminology unfortunate because these are functionals and not measures in the traditional sense. We propose to use the term Radon functional.

**Definition 5.2.** A linear functional \( \Phi: \mathcal{K}(X) \rightarrow \mathbb{C} \) is a Radon functional if for every compact subset \( K \) of \( X \), there is some real number \( c_K \geq 0 \) such that \( |\Phi(f)| \leq c_K \|f\|_{\infty} \) for all \( f \in \mathcal{K}(K; \mathbb{C}) \). The set of Radon functionals is denoted \( M(\mathcal{K}) \), or simply, \( M(X) \). The set of positive Radon functionals is denoted \( M^+(\mathcal{K}) \), and the set of continuous (or bounded) Radon functionals is denoted \( M^1(X) \).

Equivalently, a linear functional is a Radon functional if it is continuous when restricted to \( \mathcal{K}(K; \mathbb{C}) \), for every compact subset \( K \) of \( X \).

In general, a Radon functional is not continuous on \( \mathcal{K}(X) \) for the sup norm \( \| \|_{\infty} \). For a continuous Radon functional, there is a uniform constant \( c \geq 0 \) such that

\[
|\Phi(f)| \leq c \|f\|_{\infty} \quad \text{for all } f \in \mathcal{K}(X).
\]

Continuous Radon functionals are often called bounded Radon functionals.

**Proposition 5.3** immediately implies the following result.

**Proposition 5.4.** Any positive linear functional \( \Phi: \mathcal{K}(X) \rightarrow \mathbb{C} \) is a positive Radon functional.

Observe that a Radon functional \( \Phi: \mathcal{K}(X) \rightarrow \mathbb{C} \) is completely determined by its restriction \( \Phi_\mathbb{R}: \mathcal{K}(X) \rightarrow \mathbb{C} \) to the space of real-valued functions in \( \mathcal{K}(X) \). Indeed, every function \( f \in \mathcal{K}(X) \) can be written uniquely as \( f = f_1 + if_2 \) with \( f_1, f_2 \in \mathcal{K}(X) \), and by \( \mathbb{C} \)-linearity,

\[
\Phi(f) = \Phi(f_1 + if_2) = \Phi_\mathbb{R}(f_1) + i\Phi_\mathbb{R}(f_2).
\]

Also observe that \( M(X) \) and \( M^1(X) \) are vector spaces. The operator norm \( \| \| \) is well defined on the vector space \( M^1(X) \). For any bounded linear functional \( \Phi \), by definition

\[
\|\Phi\| = \sup\{|\Phi(f)| \mid f \in \mathcal{K}(X), \|f\|_{\infty} = 1\}.
\]

Using Proposition 1.9 it is easy to show that \( M^1(X) \) is isomorphic to the dual \( \mathcal{C}_0(X; \mathbb{C})' \) of the space \( \mathcal{C}_0(X; \mathbb{C}) \), that is, the space of all continuous linear forms on \( \mathcal{C}_0(X; \mathbb{C}) \). Recall that \( \mathcal{C}_0(X; \mathbb{C}) \) is the space of continuous functions which tend to 0 at infinity; see Definition 1.13.

**Proposition 5.5.** Let \( X \) be a locally compact space. The space \( M^1(X) \) of bounded Radon functionals is isomorphic to the dual \( \mathcal{C}_0(X; \mathbb{C})' \) of \( \mathcal{C}_0(X; \mathbb{C}) \), that is, the space of all continuous linear forms on \( \mathcal{C}_0(X; \mathbb{C}) \). Consequently \( M^1(X) \) is a Banach space (w.r.t. the sup norm).
Proof. By Proposition 1.9, the space $\mathcal{C}_0(X; \mathbb{C})$ is the closure of $\mathcal{K}_C(X)$. By definition, $M^1(X)$ is the space of continuous linear forms on $\mathcal{K}_C(X)$. By Theorem 13.69, every continuous linear form has a unique continuous extension to $\mathcal{C}_0(X; \mathbb{C})$. Therefore $M^1(X)$ is isomorphic to the dual of $\mathcal{C}_0(X; \mathbb{C})$. Since $\mathbb{C}$ is complete, it is known that the set of continuous linear maps from any vector space into $\mathbb{C}$ is complete. \qed

Here are some examples of Radon functionals.

Example 5.1.

1. Pick any $a \in X$. The map $\delta_a$ given by
   $$\delta_a(f) = f(a)$$
   for all $f \in \mathcal{K}_C(X)$ is a Radon functional called (with an abuse of terminology) the Dirac measure. Since $|f(a)| \leq \|f\|_{\infty}$, it is a bounded Radon functional.

2. Consider the space $\mathcal{K}_C(\mathbb{R})$ of continuous functions $f: \mathbb{R} \to \mathbb{C}$ with compact support. For each function $f \in \mathcal{K}_C(\mathbb{R})$, there is a compact interval $[a, b]$ such that $f$ vanishes outside of $[a, b]$, and from Section 2.1, the Riemann integral
   $$I(f) = \int_a^b f(t) dt$$
   is defined. We obtain a map $I: \mathcal{K}_C(\mathbb{R}) \to \mathbb{C}$ which is obviously linear. Since
   $$\left| \int_a^b f(t) dt \right| \leq (b - a) \|f\|_{\infty},$$
   this map is a Radon functional. Actually, this functional is positive. We will see later that this Radon functional corresponds to the Lebesgue measure.

3. Let $\Phi$ be any Radon functional and pick any continuous function $g \in C(X; \mathbb{C})$. It is clear that if $f \in \mathcal{K}_C(X)$, then $gf \in \mathcal{K}_C(X)$, and we have a map $\Psi$ given by
   $$\Psi(f) = \Phi(gf) \quad \text{for all } f \in \mathcal{K}_C(X).$$
   Clearly, this is a linear functional. For any compact subset $K$ of $X$, if $f \in \mathcal{K}_C(X)$, then we have
   $$\|gf\|_{\infty} \leq \|f\|_{\infty} \sup_{x \in K} |g(x)|.$$
   Since $\Phi$ is a Radon functional, there is some real $c_K \geq 0$ such that
   $$|\Phi(gf)| \leq c_K \|gf\|_{\infty},$$
   so we obtain
   $$|\Phi(gf)| \leq c_K \sup_{x \in K} |g(x)| \|f\|_{\infty},$$
which shows that $\Psi$ is a Radon functional. The Radon functional $\Psi$ is called the Radon functional with density $g$ relative to $\Phi$, and it is denoted $g \cdot \Phi$. Such Radon functionals play an important role in the definition of the notion of convolution in the theory of integration based on Radon functionals developed in Dieudonné [12] and Bourbaki [3, 5, 6, 4].

In the next section, we state the most important theorem of the theory of Radon functionals, which is that every positive Radon functional arises from a unique Borel measure with some regularity properties.

## 5.2 The Radon–Riesz Theorem and Positive Radon Functionals

In this section we deal with the direction of the correspondence positive Radon functionals $\mapsto$ Borel measures. Our first goal is to show that for every positive Radon functional $\Phi$, there is a $\sigma$-algebra $\mathcal{M}$ and a unique positive measure $m_\Phi$ on $\mathcal{M}$ (with certain properties) representing $\Phi$ as an integral, which means that

$$\Phi(f) = \int f \, dm_\Phi \quad \text{for all } f \in \mathcal{K}_C(X).$$

For instance, the positive Radon functional of Example 5.1(2) yields the Lebesgue measure. In a second stage, by imposing some reasonable conditions on the measure we obtain a bijective correspondence.

Complete proofs of these results are quite long and intricate. Such proofs can be found in Rudin [36] (Chapter 2), Lang [24] (Chapter IX) and Schwartz [41] (Chapters 5 and 7). Going back and forth between Rudin and Lang is a possible strategy to understanding the proof.

Theorem 5.6 is often referred to as the Riesz representation theorem. A version of this theorem for $X = [0, 1]$ was first proved by Frigyes Riesz¹ in 1909. In 1913, Radon extended Riesz’ result to a compact subset of $\mathbb{R}^n$ in terms of regular measures rather than a Stieltjes integral. Following Malliavin [26], it seems appropriate to call it the Radon–Riesz theorem, but it should be noted that other versions of this theorem were obtained by Banach, Saks, Markov, and Kakutani, which gives the most general version stated in Theorem 5.28; see Dunford and Schwartz [16].

**Theorem 5.6.** (Radon–Riesz) Let $X$ be a locally compact (Hausdorff) space. For every positive linear functional $\Phi : \mathcal{K}_C(X) \to \mathbb{C}$, there is a $\sigma$-algebra $\mathcal{M}$ containing the Borel $\sigma$-algebra, and there is a unique positive measure $m_\Phi$ on $\mathcal{M}$ with the following properties:

¹Not to be confused with his younger brother Marcel Riesz.
(1) The linear functional $\Phi$ is represented by $m_\Phi$, that is,
$$
\Phi(f) = \int f \, dm_\Phi \quad \text{for all } f \in K_C(X).
$$

(2) The measure $m_\Phi(K)$ is finite for every compact subset $K$ of $X$.

(3) We have
$$
m_\Phi(E) = \inf \{ m_\Phi(V) \mid E \subseteq V, \text{V open} \}
$$
for every $E \in \mathcal{M}$.

(4) We have
$$
m_\Phi(E) = \sup \{ m_\Phi(K) \mid K \subseteq E, \text{K compact} \}
$$
for every open subset $E$, and for every $E \in \mathcal{M}$ with $m_\Phi(E) < +\infty$.

(5) For any $E \in \mathcal{M}$ and any $A \subseteq E$, if $m_\Phi(E) = 0$, then $m_\Phi(A) = 0$, in other words, $m_\Phi$ is a complete measure.

Let us make a few comments about the proof. The uniqueness of $m_\Phi$ is not so bad. Observe that by (3) and (4), the measure $m_\Phi$ is determined by its values on compact subsets. Hence it suffices to prove that if two measures $\mu_1$ and $\mu_2$ satisfy the theorem, then they agree on all compact subsets.

Pick any compact $K$ and any $\epsilon > 0$. By (3) and (4), there is some open subset $V$ such that $K \subseteq V$ and $\mu_2(V) < \mu_2(K) + \epsilon$. By Proposition 13.39, there is a continuous function $f: X \to [0, 1]$ such that $f(x) = 1$ for all $x \in K$, and such that supp$(f)$ is compact and supp$(f) \subseteq V$; this implies that
$$
\mu_1(K) = \int \chi_K \, d\mu_1
\leq \int f \, d\mu_1 = \Phi(f) = \int f \, d\mu_2
\leq \int \chi_V \, d\mu_2 = \mu_2(V)
\leq \mu_2(K) + \epsilon.
$$

Therefore, $\mu_1(K) \leq \mu_2(K)$. By swapping $\mu_1$ and $\mu_2$, we obtain $\mu_2(K) \leq \mu_1(K)$, and thus $\mu_1(K) = \mu_2(K)$. Observe that the above derivation also shows that $\mu_1(K)$ is finite for every compact subset $K$.

To construct $m_\Phi$, we proceed as follows; for simplicity of notation, write $\mu$ instead of $m_\Phi$.

(a) For every open set $V$ in $X$, for every continuous function $g: X \to \mathbb{R}$, write $g < V$ if $g: X \to [0, 1]$, supp$(g)$ is compact, and supp$(g) \subseteq V$. Let
$$
\mu(V) = \sup \{ \Phi(g) \mid g < V \}.
$$
This will force Condition (4).
5.2. THE RADON–RIESZ THEOREM AND POSITIVE RADON FUNCTIONALS

(b) Next, to force Condition (3), we extend \( \mu \) to arbitrary subsets. For every \( E \subseteq X \), let

\[
\mu(E) = \inf \{ \mu(V) \mid E \subseteq V, \ V \text{ open} \}.
\]

It can be checked that \( \mu \) is an outer measure.

(c) In order to obtain a \( \sigma \)-algebra and a measure, we need to cut down the family of subsets, still forcing Conditions (3) and (4). Let \( \mathcal{A} \) be the family of all subsets \( A \) of \( X \) such that \( \mu(A) < +\infty \) and

\[
\mu(A) = \sup \{ \mu(K) \mid K \subseteq A, \ K \text{ compact} \}.
\]

Then \( \mathcal{A} \) is an algebra containing all compact sets and all open sets of finite measure. The map \( \mu \) is a measure on \( \mathcal{A} \), and if \( \mu(A) < +\infty \) then \( A \in \mathcal{A} \).

(d) Let \( \mathcal{M} \) be the family of all subsets \( Y \) of \( X \) such that \( Y \cap K \) lies in \( \mathcal{A} \) for all compact subsets \( K \). Then \( \mathcal{M} \) is the desired \( \sigma \)-algebra containing the Borel sets, and \( \mu \) is a positive measure on \( \mathcal{M} \). The algebra \( \mathcal{A} \) consists of the sets of finite measure in \( \mathcal{M} \).

Having done all this, one still needs to check that Conditions (1), (3), and (4) hold. Proposition 13.40 (existence of finite partitions of unity) is used for some of these checks.

Theorem 5.6 shows that the measure that arises from a positive linear functional has special regularity properties that we already encountered when we met the Lebesgue measure in Section 3.4.

**Definition 5.3.** A Borel measure \( \mu \) on the Borel \( \sigma \)-algebra \( \mathcal{B} \) of a locally compact space \( X \) is \( \sigma \)-regular if the following two conditions hold:

For every \( E \in \mathcal{B} \),

\[
\mu(E) = \inf \{ \mu(V) \mid E \subseteq V, \ V \text{ open} \}.
\] (**)

For every open subset \( E \), and for every \( E \in \mathcal{B} \) with \( \mu(E) < +\infty \),

\[
\mu(E) = \sup \{ \mu(K) \mid K \subseteq E, \ K \text{ compact} \}.
\] (**\(\sigma\))

Condition (**) is called outer regularity, and Condition (**\(\sigma\)) is called \( \sigma \)-inner regularity.

We say that \( \mu \) is locally finite if \( \mu(K) \) is finite for every compact subset \( K \).

The following proposition justifies the terminology \( \sigma \)-inner regularity.

**Proposition 5.7.** Let \( X \) be a locally compact (Hausdorff) space. If a Borel measure \( \mu \) is \( \sigma \)-inner regular, then

\[
\mu(E) = \sup \{ \mu(K) \mid K \subseteq E, \ K \text{ compact} \}
\] (**\(\sigma\))

holds for every \( \sigma \)-finite subset \( E \in \mathcal{B} \).
Proof. Say $E = \bigcup_{i=1}^{\infty} E_i$ with $E_i \in \mathcal{B}$ and $\mu(E_i) < +\infty$. We may assume that $\mu(E) = +\infty$, since if $\mu(E) < +\infty$ then we already have $\sigma$-inner regularity by definition. For every $M > 0$, there is some $n \geq 1$ such that $\mu\left(\bigcup_{i=1}^{n} E_i\right) > M$. Since $\bigcup_{i=1}^{n} E_i$ has finite measure, $\sigma$-inner regularity applies, so there is some compact subset $K$ such that $\mu(K) > M$. This shows that
\[ \sup\{\mu(K) \mid K \subseteq E, K \text{ compact}\} = +\infty = \mu(E), \]
which shows $\sigma$-inner regularity for $E$. \qed

Definition 5.4. Let $X$ be a locally compact (Hausdorff) space. A Borel measure $\mu$ is called a (positive) $\sigma$-Radon measure if it is $\sigma$-regular and locally finite. The space of $\sigma$-Radon measures is denoted by $\mathcal{M}_\sigma^+(X)$.

Theorem 5.6 immediately implies the following correspondence.

Theorem 5.8. (Radon–Riesz Correspondence, I) Let $X$ be a locally compact (Hausdorff) space. The maps $m: M^+(X) \to \mathcal{M}_\sigma^+(X)$ and $\varphi: \mathcal{M}_\sigma^+(X) \to M^+(X)$ given by
\[ m(\Phi) = m_\Phi \quad \text{for all } \Phi \in M^+(X) \]
\[ \varphi(\mu) = \varphi_\mu \quad \text{for all } \mu \in \mathcal{M}_\sigma^+(X) \]
are mutual inverses that define a bijection between the space $M^+(X)$ of positive Radon functionals and the space $\mathcal{M}_\sigma^+(X)$ of positive $\sigma$-Radon measures.

In the next section, we show that by requiring the locally compact space $X$ to be also $\sigma$-compact, then we obtain Borel measures that are not only $\sigma$-regular, but regular as well, which means that inner regularity holds for all $E \in \mathcal{B}$.

5.3 Regular Borel Measures

In Theorem 5.6, outer regularity holds, but $\sigma$-inner regularity holds only for open subsets and measurable sets of finite measure. It is often desirable for inner regularity to hold for arbitrary subsets $E \in \mathcal{B}$, possibly not $\sigma$-finite. It turns out that making some mild restrictions on $X$, we obtain a bijection between positive linear functionals and these regular measures. On this subject, Rudin’s exposition seems clearer than Lang’s exposition.

Definition 5.5. A Borel measure $\mu$ on the Borel $\sigma$-algebra $\mathcal{B}$ of a locally compact space $X$ is regular if the following two conditions hold for every $E \in \mathcal{B}$:
\[ \mu(E) = \inf\{\mu(V) \mid E \subseteq V, V \text{ open}\} \quad (\ast) \]
and
\[ \mu(E) = \sup\{\mu(K) \mid K \subseteq E, K \text{ compact}\} \quad (\ast\ast) \]
Condition $(\ast)$ is called outer regularity, and Condition $(\ast\ast)$ is called inner regularity.
Observe that if a Borel measure $\mu$ is $\sigma$-finite (on $X$) and if it is $\sigma$-regular, then it is actually regular. Another sufficient condition is given in Proposition 5.11.

A way to obtain the Radon–Riesz correspondence between positive Radon functionals and regular locally finite Borel measures is to require $X$ to be $\sigma$-compact, which means that $X$ is the countable union of compact subsets (see Definition 13.35).

**Theorem 5.9.** Let $X$ be a locally compact (Hausdorff), $\sigma$-compact space. For every positive linear functional $\Phi: K_C(X) \to \mathbb{C}$, if $\mathcal{M}$ and $m_\Phi$ are the $\sigma$-algebra and the measure obtained in Theorem 5.6, then the following properties holds:

1. For any $E \in \mathcal{M}$ and any $\epsilon > 0$, there is a closed set $F$ and an open set $O$ such that $F \subseteq E \subseteq O$ and $\mu(O - F) < \epsilon$.

2. The measure $m_\Phi$ is a regular, locally finite Borel measure on the Borel $\sigma$-algebra $\mathcal{B}$.

Theorem 5.9 is proved in Rudin [36] (Chapter 2, Theorem 2.17). The following theorem allows us to get a bijective correspondence between positive linear functional and regular locally finite Borel measures, and to state this theorem it is convenient to introduce the following definition.

**Definition 5.6.** Let $X$ be a locally compact (Hausdorff) space. A Borel measure $\mu$ is called a (positive) Radon measure if it is regular and locally finite. The space of Radon measures is denoted by $\mathcal{M}_\text{rad}(X)$, or simply $\mathcal{M}(X)$.

**Theorem 5.10.** (Radon–Riesz Correspondence, II) Let $X$ be a locally compact (Hausdorff), $\sigma$-compact space. The maps $m: \mathcal{M}(X) \to \mathcal{M}(X)$ and $\varphi: \mathcal{M}(X) \to \mathcal{M}(X)$ given by $m(\Phi) = m_\Phi$ for all $\Phi \in \mathcal{M}(X)$ and $\varphi(\mu) = \varphi_\mu$ for all $\mu \in \mathcal{M}(X)$ are mutual inverses that define a bijection between the space $\mathcal{M}(X)$ of positive Radon functionals and the space $\mathcal{M}(X)$ of positive Radon measures.

The following proposition gives us a sufficient condition for a locally finite Borel measure to be regular.

**Proposition 5.11.** Let $X$ be a locally compact (Hausdorff) space in which every open subset is $\sigma$-compact. If $\mu$ is a locally finite Borel measure, then $\mu$ is a regular measure.

Proposition 5.11 is proven in Rudin [36] (Chapter 2, Theorem 2.18).

Observe that $X = \mathbb{R}^n$ satisfies the condition of Proposition 5.11. Thus a locally finite Borel measure on $\mathbb{R}^n$ is a regular measure.

An interesting application of Theorem 5.10 is obtained by choosing $X = \mathbb{R}$ and $\Phi$ to be the Radon functional $I$ induced by the Riemann integral defined in Example 5.1(2). The
Radon measure $m_I$ given by Theorem 5.10 turns out to be the Lebesgue measure $\mu_L$. For details, see Rudin [36] (Chapter 2).

Measurable functions on a locally compact space with a regular, locally finite, Borel measure are very close to being continuous as stated in the following theorem of Lusin.

**Theorem 5.12.** (Lusin’s Theorem) Let $X$ be a locally compact space equipped with a regular, locally finite, Borel measure $\mu$, and let $f$ be any measurable function on $X$. If $f$ vanishes outside of a set $A$ of finite measure, for any $\epsilon > 0$, there is some function $g \in K_C(X)$ and a measurable set $Z$ with $\mu(Z) < \epsilon$, such that $f(x) = g(x)$ for all $x \in X - Z$, and $\|g\|_\infty \leq \|f\|_\infty$.

Theorem 5.12 is proven in Rudin [36] (Chapter 2, Theorem 2.24) and Lang [24] (Chapter IX, Theorem 3.3).

The Vitali–Carathéodory theorem states that every function in $L^1(X, \mathcal{B}, \mathbb{C})$ can be approximated from below and from above by certain kinds of functions called upper semicontinuous and lower semicontinuous, see Rudin [36] (Chapter 2, Theorem 2.25).

We have the following density result which uses Lusin’s theorem (Theorem 5.12).

**Theorem 5.13.** Let $X$ be a locally compact space equipped with a regular, locally finite, Borel measure $\mu$. The space $K_C(X)$ is dense in $L^p(X, \mathcal{B}, \mathbb{C})$ for $p = 1, 2$.\(^2\)

Theorem 5.13 is proved in Rudin [36] (Chapter 3, Theorem 3.14) and Lang [24] (Chapter IX, Theorem 3.1). The following proposition is needed for proving the uniqueness of the Haar measure up to a constant.

**Proposition 5.14.** Let $X$ be a locally compact space equipped with a regular, locally finite, Borel measure $\mu$. For any function $f \in L^1(X, \mathcal{B}, \mathbb{C})$, if

$$\int_A fg \, d\mu = 0 \quad \text{for all } g \in K_C(X),$$

then $f = 0$ almost everywhere.

**Proof.** We use Proposition 4.38, recalling the fact that $\int_A f \, d\mu = \int f \chi_A \, d\mu$. Let $A$ be any subset of finite measure. Then $\chi_A$ is the $L^1$-limit of a sequence $(g_n)$ of functions $g_n \in K_C(X)$ with $g_n(X) \subseteq [0, 1]$. By Proposition 4.25, there is a subsequence $(g_{n_k})_{k \geq 1}$ that converges pointwise to $\chi_A$ a.e., and thus $(fg_{n_k})$ converges pointwise to $f\chi_A$ a.e. By the dominated convergence theorem, we conclude that $\int_A f \, d\mu = \int f\chi_A \, d\mu = 0$ for all subsets $A$ of finite measure, and by Proposition 4.38, we have $f = 0$ a.e.

\(^2\)Even for all $p$ with $1 \leq p < +\infty$. 
5.4 Complex and Real Measures

By Proposition 5.2, the functionals induced by Borel measures are positive, but there are Radon functionals that are not positive, so it is natural to ask if such functionals arise from some generalized measures allowed to take negative values, or even complex values. The answer is yes. It is even possible to define measures with values in any Banach space. Such measures are discussed in Lang [24], Schwartz [41] and Marle [27], but for simplicity we will only consider real and complex measures. In this section we take a small detour to define complex measures. Then we will show how they relate to functionals on $K_C(X)$ that are not necessarily positive, but continuous.

Going back to Definition 3.9, a (positive) measure on a measurable set $(X,A)$ is a map $\mu$ satisfying the following properties:

1. $\mu: A \to [0, +\infty]$, where $A$ is a $\sigma$-algebra of subsets of $X$.
2. $\mu(\emptyset) = 0$.
3. For any countable sequence $(A_i)_{i \geq 1}$ of subsets $A_i$ of $A$ such that $A_i \cap A_j = \emptyset$ for all $i \neq j$,
   \[ \mu \left( \bigcup_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} \mu(A_i). \]

Such a function may have the value $+\infty$, but in (μ3), if $A = \bigcup_{i=1}^{\infty} A_i$ and if $\mu(A)$ is finite, then the series $\sum_{i=1}^{\infty} \mu(A_i)$ converges, and since it consists of nonnegative numbers, it converges absolutely, and thus commutatively, which means that for any permutation $\sigma$ of $\mathbb{N}_+$, we have
\[ \mu(A) = \mu \left( \bigcup_{i=1}^{\infty} A_{\sigma(i)} \right) = \sum_{i=1}^{\infty} \mu(A_{\sigma(i)}). \]

If we replace $[0, +\infty]$ by $\mathbb{R}$ or $\mathbb{C}$, then a new problem arises, namely that the convergence of the sum $\sum_{i=1}^{\infty} \mu(A_i)$ generally depends on the order of the $A_i$. The solution is to require commutative convergence of the series arising in (μ3). It is known from analysis that for $\mathbb{R}$ or $\mathbb{C}$, a series is commutatively convergent iff it is absolutely convergent, so we require the latter. We also require $\mu(A)$ be an element of $\mathbb{R}$ or $\mathbb{C}$, that is, $\mu(A)$ must be “finite.” There is a way to define measures with values in $\mathbb{R} \cup \{+\infty\}$, and even in $\mathbb{R} \cup \{-\infty, +\infty\}$, but we have no need for such generality (see Schwartz [41], Chapter V, §9).

**Definition 5.7.** Let $(X,A)$ be a measurable space. A complex measure on $(X,A)$ is a map $\mu$ satisfying the following properties:

1. $\mu: A \to \mathbb{C}$.
2. $\mu(\emptyset) = 0$. 
(μ3) For any countable family \((A_i)_{i \geq 1}\) of subsets \(A_i\) of \(\mathcal{A}\) such that \(A_i \cap A_j = \emptyset\) for all \(i \neq j\),

\[
\mu \left( \bigcup_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} \mu(A_i),
\]

where the series on the right-hand side is absolutely convergent.

A **real measure** (or **signed measure**) is a complex measure such that \(\mu(A) \subseteq \mathbb{R}\).

Observe that a real measure which is also positive is a positive measure according to Definition 3.9, but since a positive measure may take the value \(+\infty\), there are positive measures that are not real measures in the sense of Definition 5.7. When we use the term **positive real measure**, we mean that this measure only takes finite values. By **positive measure**, we mean a measure that may take the value \(+\infty\).

One might wonder if interesting real or complex measures exist. Indeed, for any arbitrary measure space \((X, \mathcal{A}, \mu)\), every function \(f \in L_\mu(X, \mathcal{A}, \mathbb{C})\) gives rise to such a measure.

**Proposition 5.15.** Let \((X, \mathcal{A}, \mu)\) be a measure space (here, \(\mu\) is a positive measure). For every integrable map \(f \in L_\mu(X, \mathcal{A}, \mathbb{C})\), the function \(\mu_f: \mathcal{A} \to \mathbb{C}\) given by

\[
\mu_f(A) = \int_A f d\mu = \int f \chi_A d\mu \quad \text{for all } A \in \mathcal{A}
\]

is a complex measure.

What is not obvious is that (μ3) holds. This follows from Proposition 4.36 (a consequence of the Lebesgue dominated convergence theorem). A detailed proof is given in Marle [27] (Chapter 1, Proposition 2.5.2).

The new twist here is that given a measure \(\mu\), rather than defining a functional by varying the function being integrated, we fix a function but we integrate by varying the subset over which we integrate.

It is trivial to check that the complex measures (and the real measure) form a vector space.

Remarkably, every complex measure \(\mu\) arises as a measure of the form \(|\mu| h\) for some suitable positive measure \(|\mu|\) and some well chosen function \(h \in L_{|\mu|}(X, \mathcal{A}, \mathbb{C})\); see Theorem 5.20. The measure \(|\mu|\) is defined as follows.

**Definition 5.8.** Let \((X, \mathcal{A})\) be a measurable space, and let \(\mu\) be a complex measure on \((X, \mathcal{A})\). Define the map \(|\mu|: \mathcal{A} \to [0, +\infty]\) by

\[
|\mu|(A) = \sup \sum_{i=1}^{\infty} |\mu(A_i)|,
\]

for all \(A \in \mathcal{A}\) and for all countable partitions \((A_i)_{i \geq 1}\) of \(A\) with \(A_i \in \mathcal{A}\). The map \(|\mu|\) is called the **total variation measure** (for short **total variation**) of \(\mu\).
5.4. COMPLEX AND REAL MEASURES

Obviously, if $\mu$ is a real positive measure, then $|\mu| = \mu$. It is easy to see that by definition,

$$|\mu(A)| \leq |\mu|(A) \quad \text{for all } A \in \mathcal{A}. $$

In fact, it is minimal with this property. We have the following remarkable theorems.

**Theorem 5.16.** Let $(X, \mathcal{A})$ be a measurable space, and let $\mu$ be a complex measure on $(X, \mathcal{A})$. The map $|\mu|: \mathcal{A} \to [0, +\infty]$ is a positive measure. The positive measure $|\mu|$ is the minimal measure such that

$$|\mu(A)| \leq |\mu|(A) \quad \text{for all } A \in \mathcal{A},$$

in the sense that if $\lambda$ is any positive positive measure such that

$$|\mu(A)| \leq \lambda(A) \quad \text{for all } A \in \mathcal{A},$$

then $|\mu| \leq \lambda$ (which means that $|\mu|(A) \leq \lambda(A)$ for all $A \in \mathcal{A}$).

A proof of Theorem 5.16 is given in Rudin[36] (Chapter 6, Theorem 6.2) and Lang [24] (Chapter VII, Theorem 3.1).

The next theorem is even more surprising.

**Theorem 5.17.** Let $(X, \mathcal{A})$ be a measurable space, and let $\mu$ be a complex measure on $(X, \mathcal{A})$. The map $|\mu|: \mathcal{A} \to [0, +\infty]$ is a finite positive measure; that is, $|\mu|(X) < +\infty$.

A proof of Theorem 5.17 is given in Rudin[36] (Chapter 6, Theorem 6.4) and Lang [24] (Chapter VII, Theorem 3.2). Theorem 5.17 implies that $\mu(X)$ is bounded: it is contained in a closed disk of finite radius. This fact shows that the convergence requirement of Condition $(\mu 3)$ is quite strong.

Theorem 5.17 allows us to make the space of complex measures into a normed vector space.

**Definition 5.9.** Let $(X, \mathcal{A})$ be measurable space. For any complex measure $\mu$, define $\|\mu\|$ as $\|\mu\| = |\mu|(X)$. The vector space of complex measures equipped with the norm defined above is denoted $\mathcal{M}_1^C(X, \mathcal{A})$, for short $\mathcal{M}_1^C(X, \mathcal{A})$.

It is not hard to show that $\mathcal{M}_1^C(X, \mathcal{A})$ is a Banach space.

**Proposition 5.18.** Let $(X, \mathcal{A})$ be a measurable space. The normed vector space $\mathcal{M}_1^C(X, \mathcal{A})$ is a Banach space (it is complete).

Another interesting fact is that if $\mu$ is a positive measure (possibly taking the value $+\infty$) then $\mathcal{L}_\mu^1(X, \mathcal{A}, \mathbb{C})$ can be embedded in $\mathcal{M}_1^C(X, \mathcal{A})$.

**Proposition 5.19.** Let $(X, \mathcal{A}, \mu)$ be a measure space. The map $f \mapsto \mu f$ is a linear embedding of $\mathcal{L}_\mu^1(X, \mathcal{A}, \mathbb{C})$ into $\mathcal{M}_1^C(X, \mathcal{A})$, and

$$\|\mu f\| = \|f\|_1 \quad \text{for all } f \in \mathcal{L}_\mu^1(X, \mathcal{A}, \mathbb{C}).$$
The next theorem shows an important fact that we mentioned earlier, namely that every complex measure $\mu$ arises as a measure of the form $|\mu|h$ for some well chosen function $h \in L_{|\mu|}(X, \mathcal{A}, \mathbb{C})$. This result is a special case of the Radon–Nikodym theorem, but for now we prefer not discussing this theorem.

**Theorem 5.20.** For every complex measure $\mu$ on a measurable space $(X, \mathcal{A})$, there is a function $h \in L_{|\mu|}(X, \mathcal{A}, \mathbb{C})$ such that $|h| = 1$ and

$$\mu(A) = \int_A h \, d|\mu| \quad \text{for all } A \in \mathcal{A}.$$ 

In other words, $\mu = |\mu|h$ (recall that $|\mu|$ is a positive measure). Furthermore, any two functions $h_1, h_2 \in L_{|\mu|}(X, \mathcal{A}, \mathbb{C})$ satisfying the conditions of the theorem are equal $|\mu|$-a.e.

For a proof of Theorem 5.20, see Rudin [36] (Chapter 6, Theorem 6.12) and Lang [24] (Chapter VII, §2 and §4).

Let us now turn our attention to real measures. We will see that any real measure can be expressed in terms of two positive real measures. This implies that any complex measure can be expressed in terms of four positive real measures. This will allow us to explain how to integrate with respect to a complex measure.

### 5.5 Real Measures and the Hahn–Jordan Decomposition

If $\mu$ is a real measure, since $|\mu|$ is a finite measure, the following real measures $\mu^+$ and $\mu^-$ are well defined.

**Definition 5.10.** If $\mu$ is a real measure, the real measures $\mu^+$ and $\mu^-$ are defined by

$$\mu^+ = \frac{1}{2}(|\mu| + \mu), \quad \mu^- = \frac{1}{2}(|\mu| - \mu).$$

It is immediately checked that $\mu^+$ and $\mu^-$ are finite positive measures, and we have

$$\mu = \mu^+ - \mu^-, \quad |\mu| = \mu^+ + \mu^-.$$ 

**Definition 5.11.** Given a real measure $\mu$, the positive real measures $\mu^+$ and $\mu^-$ are called the positive variation and negative variation of $\mu$. The expression of $\mu$ as $\mu = \mu^+ - \mu^-$ is called the Jordan decomposition of $\mu$.

The Jordan decomposition has certain minimality properties that we are going to describe.
5.5. REAL MEASURES AND THE HAHN–JORDAN DECOMPOSITION

**Definition 5.12.** Let \((X, \mathcal{A})\) be a measurable space. A complex measure \(\mu\) is concentrated on (or carried by) a measurable subset \(A\) if \(\mu(E) = 0\) for all \(E \in \mathcal{A}\) such that \(E \cap A = \emptyset\).

Two complex measures \(\mu_1\) and \(\mu_2\) are mutually singular if there exist two disjoint measurable subsets \(A_1\) and \(A_2\) such that \(\mu_1\) is supported by \(A_1\) and \(\mu_2\) is supported by \(A_2\). We sometimes write \(\mu_1 \perp \mu_2\).

Every real measure has a Hahn–Jordan decomposition as described by the following theorem.

**Theorem 5.21.** (Hahn–Jordan Decomposition) Let \((X, \mathcal{A})\) be a measurable space. For any real measure \(\mu\), there is a partition \((X^+, X^-)\) of \(X\) into two disjoint subsets of \(X\) such that

\[
\mu = \mu^+ - \mu^-
\]

is the Jordan decomposition of \(\mu\), then \(\mu^+\) is concentrated on \(X^+\), and \(\mu^-\) is concentrated on \(X^-\). Furthermore, for any \(E \in \mathcal{A}\), we have

\[
\mu^+(E) = \sup\{\mu(A) \mid A \subseteq E, A \in \mathcal{A}\}, \quad \mu^-(E) = \sup\{-\mu(A) \mid A \subseteq E, A \in \mathcal{A}\}.
\]

For any other partition \((Y^+, Y^-)\) of \(X\) such that \(\mu^+\) is concentrated on \(Y^+\) and \(\mu^-\) is concentrated on \(Y^-\),

\[
\mu^+(E \cap X^+) = \mu^+(E \cap Y^+), \quad \mu^-(E \cap X^-) = \mu^-(E \cap Y^-),
\]

for all \(E \in \mathcal{A}\).

Let us now consider a complex measure \(\mu: \mathcal{A} \to \mathbb{C}\).

**Definition 5.13.** Given a complex measure \(\mu: \mathcal{A} \to \mathbb{C}\), the function \(\overline{\mu}: \mathcal{A} \to \mathbb{C}\) called the conjugate of \(\mu\) is defined by \(\overline{\mu}(A) = \overline{\mu(A)}\) for all \(A \in \mathcal{A}\). We also define \(\mu_1: \mathcal{A} \to \mathbb{R}\) and \(\mu_2: \mathcal{A} \to \mathbb{R}\) by

\[
\mu_1(A) = \frac{1}{2}(\mu(A) + \overline{\mu(A)}), \quad \mu_2(A) = \frac{1}{2i}(\mu(A) - \overline{\mu(A)})
\]

for all \(A \in \mathcal{A}\). We call \(\mu_1\) the real part of \(\mu\) and \(\mu_2\) the imaginary part of \(\mu\).

It is immediately checked that \(\overline{\mu}\) is a complex measure, and that \(\mu_1\) and \(\mu_2\) are real measures such that

\[
\mu = \mu_1 + i\mu_2.
\]

Using the Hahn–Jordan decomposition of \(\mu_1\) an \(\mu_2\), we see that we can write \(\mu\) uniquely in terms of four positive real measure \(\mu_1^+, \mu_1^-, \mu_2^+, \mu_2^-,\) as

\[
\mu = \mu_1^+ - \mu_1^- + i(\mu_2^+ - \mu_2^-).
\]
Definition 5.14. For any complex measure $\mu : \mathcal{A} \to \mathbb{C}$, the expression

$$\mu = \mu_1^+ - \mu_2^- + i(\mu_2^+ - \mu_2^-).$$

is called the Jordan decomposition of $\mu$.

Proposition 5.22. For any complex measure $\mu : \mathcal{A} \to \mathbb{C}$, we have $|\mu_1| \leq |\mu|, |\mu_2| \leq |\mu|$, and that $|\mu| \leq |\mu_1| + |\mu_2|$. A function $f$ is $|\mu|$-integrable iff it is integrable for all four positive real measures $\mu_1^+, \mu_1^-, \mu_2^+, \mu_2^-$.

Proof. It is easy to check that $|\mu_1| \leq |\mu|, |\mu_2| \leq |\mu|$, and that $|\mu| \leq |\mu_1| + |\mu_2|$. It follows easily that $f$ is $|\mu|$-integrable if $f$ is $|\mu_1|$-integrable and $|\mu_2|$-integrable. Since $|\mu_1| = \mu_1^+ + \mu_1^-$ and $|\mu_2| = \mu_2^+ + \mu_2^-$, it is also easy to see that $f$ is $|\mu_1|$-integrable iff $f$ is $\mu_1^+$-integrable and $\mu_1^-$-integrable, and similarly $f$ is $|\mu_2|$-integrable iff $f$ is $\mu_2^+$-integrable and $\mu_2^-$-integrable. Therefore, $f$ is $|\mu|$-integrable iff it is integrable for all four positive measures $\mu_1^+, \mu_1^-, \mu_2^+, \mu_2^-$. \qed

The Jordan decomposition of the complex measure $\mu$ suggests defining the integral $\int f d\mu$ for any function $f \in L^1_{|\mu|}(X, \mathcal{A}, \mathbb{C})$.

Definition 5.15. Given any complex measure $\mu : \mathcal{A} \to \mathbb{C}$, for any function $f \in L^1_{|\mu|}(X, \mathcal{A}, \mathbb{C})$, we define the integral $\int f d\mu$ as

$$\int f d\mu = \int f d\mu_1^+ - \int f d\mu_1^- + i \int f d\mu_2^+ - i \int f d\mu_2^-.$$

By Proposition 5.22, the above expression is well defined since $f$ is $|\mu|$-integrable iff it is integrable for all four positive real measures $\mu_1^+, \mu_1^-, \mu_2^+, \mu_2^-$. Since the measures $\mu_1^+, \mu_1^-, \mu_2^+, \mu_2^-$ are positive real measures, they are finite. This immediately implies that the Radon functional $\varphi_\mu$ induced by a complex measure $\mu$ is bounded. Therefore, complex measures represent only bounded Radon functionals. Actually they represent all of them, which is the object of Section 5.7.

To show the above fact, we need to decompose a bounded Radon functional in terms of (four) positive bounded Radon functionals, and for this we introduce the notion of total variation of a Radon functional.

### 5.6 Total Variation of a Radon Functional

The notion of total variation of a Radon functional allows the decomposition of a bounded Radon functional into four positive bounded functionals, in a way that is similar to the Jordan decomposition of a complex measure. This fact is the key to the representation of a bounded Radon functional by a complex measure.

Recall that for any function $g : X \to \mathbb{C}$, we denote by $|g|$ the function $|g| : X \to \mathbb{R}$ given by $|g|(x) = |g(x)|$ for all $x \in X$.

The following result is shown in Dieudonné [12] (Chapter XIII, Section 3).
Theorem 5.23. For any Radon functional $\Phi: K_C(X) \to \mathbb{C}$ on a locally compact space $X$, there is a smallest positive Radon functional $|\Phi|: K_C(X) \to \mathbb{C}$ such that

$$|\Phi(f)| \leq |\Phi|(|f|) \quad \text{for all } f \in K_C(X).$$

The functional $|\Phi|$ is completely defined by its restriction to positive functions $f \geq 0$ in $K_{\mathbb{R}}(X)$ by

$$|\Phi|(f) = \sup\{\Phi(g) \mid g \in K_C(X), |g| \leq f\}.$$ 

Definition 5.16. Given any Radon functional $\Phi: K_C(X) \to \mathbb{C}$, the positive Radon functional $|\Phi|$ is called total variation (or absolute value) of $\Phi$.

If $\Phi$ is a positive Radon functional, then

$$|\Phi| = \Phi.$$ 

Definition 5.17. Given a Radon functional $\Phi: K_C(X) \to \mathbb{C}$, we define the conjugate $\overline{\Phi}$ of $\Phi$ by

$$\overline{\Phi}(f) = \overline{\Phi(f)}, \quad f \in K_C(X).$$

If we write $f = f_1 + if_2$ with $f_1, f_2 \in K_{\mathbb{R}}(X)$, then we have

$$\overline{\Phi}(f) = \overline{\Phi}(f_1 + if_2) = \overline{\Phi((f_1 + if_2))} = \overline{\Phi(f_1 - if_2)} = (\overline{\Phi(f_1)} - i\Phi(f_2)) = \overline{\Phi(f_1)} + i\Phi(f_2).$$

Definition 5.18. We say that a Radon functional $\Phi: K_C(X) \to \mathbb{C}$ is real if $\overline{\Phi} = \Phi$.

Proposition 5.24. A Radon functional $\Phi: K_C(X) \to \mathbb{C}$ is real iff its restriction $\Phi_{\mathbb{R}}$ to $K_{\mathbb{R}}(X)$ is a real-valued function $\Phi_{\mathbb{R}}: K_{\mathbb{R}}(X) \to \mathbb{R}$.

Proof. In view of the above computation, a Radon functional $\Phi$ is real iff

$$\Phi(f_1) + i\Phi(f_2) = \overline{\Phi(f_1)} + i\overline{\Phi(f_2)}$$

for all $f_1, f_2 \in K_{\mathbb{R}}(X)$, which by setting $f_2 = 0$ or $f_1 = 0$ means that $\Phi(f_1) \in \mathbb{R}$ for all $f_1 \in K_{\mathbb{R}}(X)$. Equivalently, a Radon functional $\Phi: K_C(X) \to \mathbb{C}$ is real iff its restriction $\Phi_{\mathbb{R}}$ to $K_{\mathbb{R}}(X)$ is a real-valued function $\Phi_{\mathbb{R}}: K_{\mathbb{R}}(X) \to \mathbb{R}$. 

Since a Radon functional is $\Phi$ is completely determined by its restriction $\Phi_{\mathbb{R}}$ to $K_{\mathbb{R}}(X)$, we often think of a real Radon functional as a linear map $\Phi: K_{\mathbb{R}}(X) \to \mathbb{R}$.

Definition 5.19. Given a Radon functional $\Phi: K_C(X) \to \mathbb{C}$, we define $\Phi_r$ and $\Phi_i$ by

$$\Phi_r = \frac{1}{2}(\Phi + \overline{\Phi}), \quad \Phi_i = \frac{1}{2i}(\Phi - \overline{\Phi}).$$
It is immediately verified that $\Phi_r$ and $\Phi_i$ are real Radon functionals such that

$$\Phi = \Phi_r + i\Phi_i, \quad \Phi = \Phi_r - i\Phi_i.$$ \[\text{We also have}\]

$$|\Phi_r| \leq |\Phi|, \quad |\Phi_i| \leq |\Phi|, \quad |\Phi| \leq |\Phi_r| + |\Phi_i|.$$

**Definition 5.20.** If $\Phi: K_C(X) \to \mathbb{R}$ is a real Radon functional, then as in the case of real measures we can define $\Phi^+$ and $\Phi^-$ by

$$\Phi^+ = \frac{1}{2}(|\Phi| + \Phi), \quad \Phi^- = \frac{1}{2}(|\Phi| - \Phi).$$

It is immediately checked that $\Phi^+$ and $\Phi^-$ are positive Radon functionals, and we have

$$\Phi = \Phi^+ - \Phi^-, \quad |\Phi| = \Phi^+ + \Phi^-.$$

In the end, we have the following decomposition result analogous to the Jordan decomposition for complex measures.

**Proposition 5.25.** Every Radon functional $\Phi: K_C(X) \to \mathbb{C}$ can be expressed in terms of four positive Radon functionals:

$$\Phi = \Phi^+_r - \Phi^-_r + i(\Phi^+_i - \Phi^-_i).$$

By the Radon–Riesz theorem (Theorem 5.6), there exist four positive measures $m_1, m_2, m_3, m_4$ such that

$$\Phi(f) = \int f \, dm_1 - \int f \, dm_2 + i \left( \int f \, dm_3 - \int f \, dm_4 \right) \quad \text{for all } f \in K_C(X).$$

It is tempting to define the complex measure $m$ by

$$m = m_1 - m_2 + i(m_3 - m_4),$$

but there is a problem, which is that the positive measures $m_i$ may take the value $+\infty$, so expressions of the form $+\infty - (+\infty)$ may arise, but they do not make any sense!

We are not aware of a way around this problem in general. If $X$ is compact, then the Radon–Riesz theorem yields Borel measures $m_i$ such that $m_i(X)$ is finite for $i = 1, \ldots, 4$, in which case the expression $m$ is indeed a measure. It is even possible to define a bijective correspondence by adding disjointness conditions on the subsets over which the $m_i$ are concentrated. Such results are given in Malliavin [26] (Chapter II, Section 5).

Another situation where $m$ is a complex measure is the case where the Radon functional $\Phi$ is bounded (continuous). This is the object of the next section.
5.7 The Radon–Riesz Theorem and Bounded Radon Functionals

Let $\Phi: \mathcal{K}_C(X) \to \mathbb{C}$ be a bounded Radon functional. In this case the operator norm $\|\Phi\|$ is finite. Recall that
\[
\|\Phi\| = \sup\{\|\Phi(f)\| : f \in \mathcal{K}_C(X), \|f\|_\infty \leq 1\} = \sup\{\|\Phi(f)\| : f \in \mathcal{K}_C(X), \|f\|_\infty = 1\}.
\]

The following result is shown in Dieudonné [12] (Chapter VII, Section 20).

**Proposition 5.26.** Given a Radon functional $\Phi: \mathcal{K}_C(X) \to \mathbb{C}$, the norm $\|\Phi\|$ is finite, that is, $\Phi$ is bounded, iff $|\Phi|$ is bounded. In this case, $\|\Phi\| = \||\Phi||$.

We deduce that $\Phi$ is bounded iff $\Phi_r$ and $\Phi_i$ are bounded. But we also see that a real bounded Radon functional $\Psi$ is bounded iff the positive Radon functionals $\Psi^+$ and $\Psi^-$ are bounded.

**Proposition 5.27.** A Radon functional $\Phi$ is bounded iff the positive Radon functional $\Phi^+_r, \Phi^+_i, \Phi^-_r, \Phi^-_i$ are bounded.

If $m_1, m_2, m_3, m_4$ are the positive Borel measures representing $\Phi^+_r, \Phi^-_r, \Phi^+_i, \Phi^-_i$ given by the Radon–Riesz theorem (Theorem 5.6), it turns out that they are all finite measures, so $m = m_1 - m_2 + i(m_3 - m_4)$ is a complex measure, and it represents $\Phi$ on functions in $\mathcal{C}_0(X)$. In order to state a suitable version of the Radon–Riesz correspondence, we need the following definition.

**Definition 5.21.** Let $X$ be a locally compact (Hausdorff) space. A complex measure $\mu$ on the $\sigma$-algebra $\mathcal{B}$ of Borel sets of $X$ is a regular complex Borel measure if the positive measure $|\mu|$ is regular and locally finite (that is, $|\mu|(K)$ is finite for every compact subset $K$). We denote the vector space of regular complex Borel measures by $\mathcal{M}^{1}_{\text{reg}, \mathbb{C}}(X)$.

We have the following beautiful theorem. Theorem 5.28 is also often referred to as the Riesz representation theorem, which is somewhat confusing.

**Theorem 5.28.** (Radon–Riesz Correspondence, III) Let $X$ be a locally compact (Hausdorff) space. There are bijections $m: \mathcal{M}^1(X) \to \mathcal{M}^{1}_{\text{reg}, \mathbb{C}}(X)$ and $\varphi: \mathcal{M}^{1}_{\text{reg}, \mathbb{C}}(X) \to \mathcal{M}^1(X)$ between the Banach space $\mathcal{M}^1(X) = \mathcal{C}_0(X, \mathbb{C})'$ of bounded Radon functionals, the dual of the space $\mathcal{C}_0(X, \mathbb{C})$ of continuous functions that tend to zero at infinity, and the Banach space $\mathcal{M}^{1}_{\text{reg}, \mathbb{C}}(X)$ of regular complex regular Borel measures. For every bounded Radon functional $\Phi \in \mathcal{C}_0(X, \mathbb{C})'$, the complex regular Borel measure $m_\Phi$ represents $\Phi$ in the sense that
\[
\Phi(f) = \int f dm_\Phi
\]
\[
= \int f dm_\Phi^+_r - \int f dm_\Phi^-_r + i \left( \int f dm_\Phi^+_i - \int f dm_\Phi^-_i \right) \quad \text{for all } f \in \mathcal{C}_0(X, \mathbb{C}).
\]

Furthermore, these bijections are norm preserving, that is, $\|\Phi\| = \|m_\Phi\| = |m_\Phi|(X)$. 
Theorem 5.28 is proven in Lang [24] (Chapter IX, §4, Theorem 4.2), Rudin [36] (Chapter 6 Theorem 6.19), and Marle [27] (Chapter 9, Section 7, Proposition 9.7.3). The proof is quite involved. Among other things it uses Lusin’s theorem (Theorem 5.12) and the corollary of the Radon–Nikodym theorem (Theorem 5.20).

To prove surjectivity, by Proposition 5.25 we express the bounded Radon functional $\Phi: \mathcal{K}_C(X) \to \mathbb{C}$ in terms of four positive Radon functionals:

$$\Phi = \Phi_+^r - \Phi_-^r + i(\Phi_+^i - \Phi_-^i).$$

By Proposition 5.27, these positive Radon functionals are bounded. By the Radon–Riesz theorem (Theorem 5.6), there exist four positive $\sigma$-regular Borel measures $m_1, m_2, m_3, m_4$ such that

$$\Phi(f) = \int f dm_1 - \int f dm_2 + i \left( \int f dm_3 - \int f dm_4 \right)$$

for all $f \in \mathcal{K}_C(X)$.

The reason why the Borel measure $m$ corresponding to a positive bounded Radon functional $\Phi$ is finite is that this measure is inner regular, that is,

$$m_\Phi(E) = \sup\{\mu(K) \mid K \subseteq E, K \text{ compact}\}$$

for every $E \in \mathcal{B}$. We use this to compute $m_\Phi(X)$. For every compact subset $K$, by Proposition 13.39, there is a continuous function $f : X \to [0, 1]$ of compact support such that $f(x) = 0$ for all $x \in K$. Then since $\Phi$ is bounded we have

$$m_\Phi(K) \leq \int f dm_\Phi = \Phi(f) \leq \|\Phi\| \|f\|_\infty = \|\Phi\|$$

since $f$ has maximum value 1. Therefore,

$$m_\Phi(X) = \sup\{\mu(K) \mid K \subseteq X, K \text{ compact}\} \leq \|\Phi\|$$

is indeed finite. Since $m_\Phi(X)$ is finite, the $\sigma$-regular measure $m_\Phi$ is actually regular.

We also need to check that $\int f dm$ is finite for every function $f \in \mathcal{C}_0(X; \mathbb{C})$ and every positive finite Borel measure $m$. Since $\mathcal{C}_0(X; \mathbb{C})$ is the closure of $\mathcal{K}_C(X)$, there is a sequence $(f_n)$ of functions $f_n \in \mathcal{K}_C(X)$ that converges to $f$ according to the sup norm, and thus converges pointwise to $f$. Also $f$ is a bounded function, so there is some $M > 0$ such that $|f_n| \leq M$ for all $n \geq 1$. Since $m(X)$ is finite, the constant function $M$ is integrable, and the continuous functions $f_n$ are integrable. By the dominated convergence theorem (Theorem 4.33), $f$ is integrable.

Theorem 5.28 plays a crucial role in defining the notion of convolutions of two measures in $\mathcal{M}^1_{\text{reg}, \mathbb{C}}(X)$. We will need the following simple fact.
Proposition 5.29. Let \( X \) be any locally compact space, and let \( \mu \) be any positive Borel measure on \( B \). For any function \( f \in L^1_\mu(X, B, \mathbb{C}) \), the functional \( \Phi_{f,\mu} : C_0(X; \mathbb{C}) \rightarrow \mathbb{C} \) given by
\[
\Phi_{f,\mu}(g) = \int fg \, d\mu \quad \text{for all } g \in C_0(X; \mathbb{C})
\]
is a bounded Radon functional.

Proof. Since \( f \in L^1_\mu(X, B, \mathbb{C}) \) and \( g \) is continuous, \( g \) is measurable, and \( |g| \) is bounded by some \( M > 0 \), so by Proposition 4.35(1) \( fg \in L^1_\mu(X, B, \mathbb{C}) \). We have
\[
|\Phi_{f,\mu}(g)| = \left| \int fg \, d\mu \right| \leq \int |fg| \, d\mu = \int |f||g| \, d\mu \leq \|g\|_\infty \int |f| \, d\mu,
\]
which shows that \( \Phi_{f,\mu} \) is bounded. \( \square \)

By Theorem 5.28, the bounded Radon functional \( \Phi_{f,\mu} \) corresponds to a unique complex regular measure \( m \) such that
\[
\int fg \, d\mu = \int g \, dm \quad \text{for all } g \in C_0(X; \mathbb{C}).
\]
The measure \( m \) is usually denoted by \( fd\mu \); for example, see Folland [18] (Chapter 2, Section 2.5). Proposition 5.29 gives us an embedding of \( L^1_\mu(X, B, \mathbb{C}) \) into \( M^1_{\text{reg},\mathbb{C}}(X) \) as stated in the next proposition.

Proposition 5.30. Let \( X \) be a locally compact space. For every positive Borel measure \( \mu \) on \( B \), the map \( f \mapsto fd\mu \) is an embedding of \( L^1_\mu(X, B, \mathbb{C}) \) into \( M^1_{\text{reg},\mathbb{C}}(X) \) of complex regular Borel measures on \( X \), with the property that
\[
\int fg \, d\mu = \int g \, fd\mu \quad \text{for all } g \in C_0(X; \mathbb{C}).
\]

One of the reasons for using this embedding is that if \( X \) is a locally compact group and \( \mu \) is a Haar measure, convolution can be defined on both \( L^1_\mu(X, B, \mathbb{C}) \) and \( M^1_{\text{reg},\mathbb{C}}(X) \), but there is no identity element for convolution on \( L^1_\mu(X, B, \mathbb{C}) \) while there is one for convolution on \( M^1_{\text{reg},\mathbb{C}}(X) \).
Chapter 6

The Haar Measure

6.1 Topological Groups

Since locally compact groups (and Lie groups) are topological groups, it is useful to gather a few basic facts about topological groups.

Definition 6.1. A set $G$ is a topological group iff

(a) $G$ is a Hausdorff topological space;

(b) $G$ is a group (with identity $1$);

(c) Multiplication $\cdot: G \times G \to G$, and the inverse operation $G \to G: g \mapsto g^{-1}$, are continuous, where $G \times G$ has the product topology.

It is easy to see that the two requirements of Condition (c) are equivalent to

(c') The map $G \times G \to G: (g, h) \mapsto gh^{-1}$ is continuous.

Proposition 6.1. If $G$ is a topological group and $H$ is any subgroup of $G$, then the closure $\overline{H}$ of $H$ is a subgroup of $G$. If $H$ is a normal subgroup of $G$, then $\overline{H}$ is also a normal subgroup of $G$.

Proof. We use the fact that if $f: X \to Y$ is a continuous map between two topological spaces $X$ and $Y$, then $f(A) \subseteq \overline{f(A)}$ for any subset $A$ of $X$. For any $a \in \overline{A}$, we need to show that for any open subset $W \subseteq Y$ containing $f(a)$, we have $W \cap f(A) \neq \emptyset$. Since $f$ is continuous, $V = f^{-1}(W)$ is an open subset containing $a$, and since $a \in \overline{A}$, we have $f^{-1}(W) \cap A \neq \emptyset$, so there is some $x \in f^{-1}(W) \cap A$, which implies that $f(x) \in W \cap f(A)$, so $W \cap f(A) \neq \emptyset$, as desired. The map $f: G \times G \to G$ given by $f(x, y) = xy^{-1}$ is continuous, and since $H$ is a subgroup of $G$, $f(H \times H) \subseteq H$. By the above property, if $a \in \overline{H}$ and if $b \in \overline{H}$, that is, $(a, b) \in \overline{H} \times \overline{H}$, then $f(a, b) = ab^{-1} \in \overline{H}$, which shows that $\overline{H}$ is a subgroup of $G$.

For every $g \in G$, the map $C_g: G \to G$ given by $C_g(x) = gxg^{-1}$ for all $x \in G$ is continuous, and if $H$ is a normal subgroup of $G$, then $C_g(H) \subseteq H$. It follows that $C_g(\overline{H}) \subseteq \overline{H}$ for all $g \in G$, which means that $\overline{H}$ is a normal subgroup of $G$. \hfill \square
Given a topological group $G$, for every $a \in G$ we define the left translation $L_a$ as the map $L_a : G \to G$ such that $L_a(b) = ab$, for all $b \in G$, and the right translation $R_a$ as the map $R_a : G \to G$ such that $R_a(b) = ba$, for all $b \in G$. Observe that $L_a^{-1}$ is the inverse of $L_a$ and similarly, $R_a^{-1}$ is the inverse of $R_a$. As multiplication is continuous, we see that $L_a$ and $R_a$ are continuous. Moreover, since they have a continuous inverse, they are homeomorphisms. As a consequence, if $U$ is an open subset of $G$, then so is $gU = L_g(U)$ (resp. $Ug = R_gU$), for all $g \in G$. Therefore, the topology of a topological group is determined by the knowledge of the open subsets containing the identity 1.

Given any subset $S \subseteq G$, let $S^{-1} = \{ s^{-1} \mid s \in S \}$; let $S^0 = \{1\}$, and $S^{n+1} = S^n S$, for all $n \geq 0$. Property (c) of Definition 6.1 has the following useful consequences, which shows there exists an open subset containing 1 which has a special symmetrical structure.

**Proposition 6.2.** If $G$ is a topological group and $U$ is any open subset containing 1, then there is some open subset $V \subseteq U$, with $1 \in V$, so that $V = V^{-1}$ and $V^2 \subseteq U$. Furthermore, $\overline{V} \subseteq U$.

**Proof.** Since multiplication $G \times G \to G$ is continuous and $G \times G$ is given the product topology, there are open subsets $U_1$ and $U_2$, with $1 \in U_1$ and $1 \in U_2$, so that $U_1 U_2 \subseteq U$. Let $W = U_1 \cap U_2$ and $V = W \cap W^{-1}$. Then $V$ is an open set containing 1, and clearly $V = V^{-1}$ and $V^2 \subseteq U$. If $g \in \overline{V}$, then $gV$ is an open set containing $g$ (since $1 \in V$) and thus, $gV \cap V \neq \emptyset$. This means that there are some $h_1, h_2 \in V$ so that $gh_1 = h_2$, but then, $g = h_2 h_1^{-1} \in VV^{-1} = V^2 \subseteq U$. \hfill \Box

**Definition 6.2.** A subset $U$ containing 1 and such that $U = U^{-1}$ is called symmetric.

Proposition 6.2 is used in the proofs of many of the propositions and theorems on the structure of topological groups. For example, it is key in verifying the following proposition regarding discrete topological subgroups.

**Definition 6.3.** A subgroup $H$ of a topological group $G$ is discrete iff the induced topology on $H$ is discrete; that is, for every $h \in H$, there is some open subset $U$ of $G$ so that $U \cap H = \{h\}$.

**Proposition 6.3.** If $G$ is a topological group and $H$ is a discrete subgroup of $G$, then $H$ is closed.

**Proof.** As $H$ is discrete, there is an open subset $U$ of $G$ so that $U \cap H = \{1\}$, and by Proposition 6.2, we may assume that $U = U^{-1}$. Our goal is to show $H = \overline{H}$. Clearly $H \subseteq \overline{H}$. Thus it remains to show $\overline{H} \subseteq H$. If $g \in \overline{H}$, as $gU$ is an open set containing $g$, we have $gU \cap H \neq \emptyset$. Consequently, there is some $y \in gU \cap H = gU^{-1} \cap H$, so $g \in yU$ with $y \in H$. We claim that $yU \cap H = \{y\}$. Note that $x \in yU \cap H$ means $x = yu_1$ with $yu_1 \in H$ and $u_1 \in U$. Since $H$ is a subgroup of $G$ and $y \in H$, $y^{-1}yu_1 = u_1 \in H$. Thus $u_1 \in U \cap H$, which implies $u_1 = 1$ and $x = yu_1 = y$, and we have $g \in yU \cap H \subseteq \overline{yU \cap H} = \overline{\{y\}} = \{y\}$. since $G$ is Hausdorff. Therefore, $g = y \in H$. \hfill \Box
Using Proposition 6.2, we can give a very convenient characterization of the Hausdorff separation property in a topological group.

**Proposition 6.4.** If $G$ is a topological group, then the following properties are equivalent:

1. $G$ is Hausdorff;
2. The set $\{1\}$ is closed;
3. The set $\{g\}$ is closed, for every $g \in G$.

**Proof.** The implication (1) $\rightarrow$ (2) is true in any Hausdorff topological space. We just have to prove that $G - \{1\}$ is open, which goes as follows: For any $g \neq 1$, since $G$ is Hausdorff, there exists disjoint open subsets $U_g$ and $V_g$, with $g \in U_g$ and $1 \in V_g$. Thus, $\bigcup U_g = G - \{1\}$, showing that $G - \{1\}$ is open. Since $L_g$ is a homeomorphism, (2) and (3) are equivalent.

Let us prove that (3) $\rightarrow$ (1). Let $g_1, g_2 \in G$ with $g_1 \neq g_2$. Then, $g_1^{-1}g_2 \neq 1$ and if $U$ and $V$ are disjoint open subsets such that $1 \in U$ and $g_1^{-1}g_2 \in V$, then $g_1 \in g_1U$ and $g_2 \in g_1V$, where $g_1U$ and $g_1V$ are still open and disjoint. Thus, it is enough to separate 1 and $g \neq 1$. Pick any $g \neq 1$. If every open subset containing 1 also contained $g$, then 1 would be in the closure of $\{g\}$, which is absurd since $\{g\}$ is closed and $g \neq 1$. Therefore, there is some open subset $U$ such that $1 \in U$ and $g \notin U$. By Proposition 6.2, we can find an open subset $V$ containing 1, so that $VV \subseteq U$ and $V = V^{-1}$. We claim that $V$ and $gV$ are disjoint open sets with $1 \in V$ and $g \in gV$.

Since $1 \in V$, it is clear that $g \in gV$. If we had $V \cap gV \neq \emptyset$, then by the last sentence in the proof of Proposition 6.2, we would have $g \in VV^{-1} = VV \subseteq U$, a contradiction. 

If $H$ is a subgroup of $G$ (not necessarily normal), we can form the set of left cosets $G/H$, and we have the projection $p: G \rightarrow G/H$, where $p(g) = gH = \overline{g}$. If $G$ is a topological group, then $G/H$ can be given the quotient topology, where a subset $U \subseteq G/H$ is open iff $p^{-1}(U)$ is open in $G$. With this topology, $p$ is continuous. The trouble is that $G/H$ is not necessarily Hausdorff. However, we can neatly characterize when this happens.

**Proposition 6.5.** If $G$ is a topological group and $H$ is a subgroup of $G$, then the following properties hold:

1. The map $p: G \rightarrow G/H$ is an open map, which means that $p(V)$ is open in $G/H$ whenever $V$ is open in $G$.
2. The space $G/H$ is Hausdorff iff $H$ is closed in $G$.
3. If $H$ is open, then $H$ is closed and $G/H$ has the discrete topology (every subset is open).
4. The subgroup $H$ is open iff $1 \in \overset{\circ}{H}$ (i.e., there is some open subset $U$ so that $1 \in U \subseteq H$).
Proof. (1) Observe that if $V$ is open in $G$, then $VH = \bigcup_{h \in H} Vh$ is open, since each $Vh$ is open (as right translation is a homeomorphism). However, it is clear that

$$p^{-1}(p(V)) = VH,$$

i.e., $p^{-1}(p(V))$ is open which, by definition of the quotient topology, means that $p(V)$ is open.

(2) If $G/H$ is Hausdorff, then by Proposition 6.4, every point of $G/H$ is closed, i.e., each coset $gH$ is closed, so $H$ is closed. Conversely, assume $H$ is closed. Let $\overline{x}$ and $\overline{y}$ be two distinct points in $G/H$ and let $x, y \in G$ be some elements with $p(x) = \overline{x}$ and $p(y) = \overline{y}$. As $\overline{x} \neq \overline{y}$, the elements $x$ and $y$ are not in the same coset, so $x \notin yH$. As $H$ is closed, so is $yH$, and since $x \notin yH$, there is some open containing $x$ which is disjoint from $yH$, and we may assume (by translation) that it is of the form $Ux$, where $U$ is an open containing $1$. By Proposition 6.2, there is some open $V$ containing $1$ so that $VV \subseteq U$ and $V = V^{-1}$. Thus, we have

$$V^2x \cap yH = \emptyset$$

and in fact,

$$V^2xH \cap yH = \emptyset,$$

since $H$ is a group; if $z \in V^2xH \cap yH$, then $z = v_1v_2xh_1 = yh_2$ for some $v_1, v_2 \in V$, and some $h_1, h_2 \in H$, but then $v_1v_2 = yh_2h_1^{-1}$ so that $V^2x \cap yH \neq \emptyset$, a contradiction. Since $V = V^{-1}$, we get

$$VxH \cap VyH = \emptyset,$$

and then, since $V$ is open, both $VxH$ and $VyH$ are disjoint, open, so $p(VxH)$ and $p(VyH)$ are open sets (by (1)) containing $\overline{x}$ and $\overline{y}$ respectively and $p(VxH)$ and $p(VyH)$ are disjoint (because $p^{-1}(p(VxH)) = VxHH = VxH$, $p^{-1}(p(VyH)) = VyHH = VyH$, and $VxH \cap VyH = \emptyset$). See Figure 6.1.

(3) If $H$ is open, then every coset $gH$ is open, so every point of $G/H$ is open and $G/H$ is discrete. Also, $\bigcup_{g \notin H} gH$ is open, i.e., $H$ is closed.

(4) Say $U$ is an open subset such that $1 \in U \subseteq H$. Then for every $h \in H$, the set $hU$ is an open subset of $H$ with $h \in hU$, which shows that $H$ is open. The converse is trivial. \qed

Recall that a topological space $X$ is locally compact iff for every point $p \in X$, there is a compact neighborhood $C$ of $p$; that is, there is a compact $C$ and an open $U$, with $p \in U \subseteq C$. For example, manifolds are locally compact.

The next two propositions will be needed.

**Proposition 6.6.** Let $G$ be a topological group and let $H$ be a closed subgroup of $G$. The following properties hold.

(1) If $G$ is locally compact, then so is $G/H$.  

(2) If $H$ is a normal subgroup of $G$, then $G/H$ is a topological group.

Proof. (1) Since $H$ is closed, we already know from Proposition 6.5(2) that $G/H$ is Hausdorff. Let $K$ be a compact neighborhood of 1 in $G$, so that there is an open subset $U$ such that $1 \in U \subseteq K$ with $K$ compact. By Proposition 6.5(1) the quotient map $p: G \to G/H$ is an open map, and it is continuous, so for any $g \in G$, we have $g \in gU \subseteq gK$ with $gU$ open and $gK$ compact, so $p(g) \in p(gU) \subseteq p(gK)$ with $p(gU)$ open and $(gK)$ compact, which shows that $G/H$ is locally compact.

(2) If $H$ is a closed normal subgroup, then $G/H$ is a group, and we already know from Proposition 6.5(2) that $G/H$ is Hausdorff. We have to show that multiplication and inversion in $G/H$ are continuous. For any two cosets $g_1H$ and $g_2H$ in $G/H$, if $W$ is an open subset in $G/H$ containing $p(g_1g_2) = p(g_1)p(g_2) = (g_1H)(g_2H) = g_1g_2H$, then because the projection map $p$ is continuous, there are open subsets $U_1$ and $U_2$ of $G$ with $g_1 \in U_1$ and $g_2 \in U_2$, such that $p(U_1U_2) \subseteq W$. Since $p$ is an open map, $p(U_1)$ is an open subset containing $p(g_1) = g_1H$ and $p(U_2)$ is an open subset containing $p(g_2) = g_2H$, and we have $p(U_1)p(U_2) \subseteq W$, so multiplication in $G/H$ is continuous. A similar proof shows that inversion is continuous in $G/H$.

Proposition 6.7. Let $G$ be a locally compact topological group and let $H$ be a closed subgroup of $G$. For any compact subset $K'$ in $G/H$, there is compact subset $K$ of $G$ such that $p(K) = K'$.

Proof. Since $G$ is locally compact, there is an open subset $U$ and a compact subset $V$ such that $1 \in U \subseteq V$. Since $p$ is an open map, the subsets of the form $p(gU)$ for $g \in G$ form an open cover of $K'$, and since $K'$ is compact, there is a finite subcover $\{p(g_1U), \ldots, p(g_nU)\}$ of $K'$. Since $p$ is continuous and $K'$ is compact and thus closed (since $G/H$ is Hausdorff), $p^{-1}(K')$ is closed and $g_1V \cup \cdots \cup g_nV$ is compact, $K = p^{-1}(K') \cap (g_1V \cup \cdots \cup g_nV)$ is compact in $G$, and we have $p(K) = K'$.
We next provide a criterion relating the connectivity of \( G \) with that of \( G/H \).

**Proposition 6.8.** Let \( G \) be a topological group and \( H \) be any subgroup of \( G \). If \( H \) and \( G/H \) are connected, then \( G \) is connected.

**Proof.** It is a standard fact of topology that a space \( G \) is connected iff every continuous function \( f \) from \( G \) to the discrete space \( \{0,1\} \) is constant; see Proposition 13.17. Pick any continuous function \( f \) from \( G \) to \( \{0,1\} \). As \( H \) is connected and left translations are homeomorphisms, all cosets \( gH \) are connected. Thus, \( f \) is constant on every coset \( gH \). It follows that the function \( f : G \to \{0,1\} \) induces a continuous function \( \overline{f} : G/H \to \{0,1\} \) such that \( f = \overline{f} \circ p \) (where \( p : G \to G/H \); the continuity of \( \overline{f} \) follows immediately from the definition of the quotient topology on \( G/H \)). As \( G/H \) is connected, \( \overline{f} \) is constant, and so \( f = \overline{f} \circ p \) is constant. \( \square \)

The next three propositions describe how to generate a topological group from its symmetric neighborhoods of 1.

**Proposition 6.9.** If \( G \) is a connected topological group, then \( G \) is generated by any symmetric neighborhood \( V \) of 1. In fact,

\[
G = \bigcup_{n \geq 1} V^n.
\]

**Proof.** Since \( V = V^{-1} \), it is immediately checked that \( H = \bigcup_{n \geq 1} V^n \) is the group generated by \( V \). As \( V \) is a neighborhood of 1, there is some open subset \( U \subseteq V \), with \( 1 \in U \), and so \( 1 \in H \). From Proposition 6.5 (3), the subgroup \( H \) is open and closed, and since \( G \) is connected, \( H = G \). \( \square \)

**Proposition 6.10.** Let \( G \) be a topological group and let \( V \) be any connected symmetric open subset containing 1. Then, if \( G_0 \) is the connected component of the identity, we have

\[
G_0 = \bigcup_{n \geq 1} V^n,
\]

and \( G_0 \) is a normal subgroup of \( G \). Moreover, the group \( G/G_0 \) is discrete.

**Proof.** First, as \( V \) is open, every \( V^n \) is open, so the group \( \bigcup_{n \geq 1} V^n \) is open, and thus closed, by Proposition 6.5 (3). For every \( n \geq 1 \), we have the continuous map

\[
V \times \cdots \times V \longrightarrow V^n : (g_1, \ldots, g_n) \mapsto g_1 \cdots g_n.
\]

As \( V \) is connected, \( V \times \cdots \times V \) is connected, and so \( V^n \) is connected; this follows from Proposition 13.18 because a finite product of connected spaces is connected. Since \( 1 \in V^n \) for all \( n \geq 1 \) and every \( V^n \) is connected, we use Lemma 13.19 to conclude that \( \bigcup_{n \geq 1} V^n \) is connected. Now, \( \bigcup_{n \geq 1} V^n \) is connected, open and closed, so it is the connected component of 1. Finally, for every \( g \in G \), the group \( gG_0g^{-1} \) is connected and contains 1, so it is contained in \( G_0 \), which proves that \( G_0 \) is normal. Since \( G_0 \) is open, Proposition 6.5 (3) implies that the group \( G/G_0 \) is discrete. \( \square \)
Proposition 6.11. Let $G$ be a topological group and assume that $G$ is connected and locally compact. Then, $G$ is countable at infinity, which means that $G$ is the union of a countable family of compact subsets. In fact, if $V$ is any symmetric compact neighborhood of 1, then

$$G = \bigcup_{n \geq 1} V^n.$$ 

Proof. Since $G$ is locally compact, there is some compact neighborhood $K$ of 1. Then, $V = K \cap K^{-1}$ is also compact and a symmetric neighborhood of 1. By Proposition 6.9, we have

$$G = \bigcup_{n \geq 1} V^n.$$ 

An argument similar to the one used in the proof of Proposition 6.10 to show that $V^n$ is connected if $V$ is connected proves that each $V^n$ compact if $V$ is compact.

If $G$ is a locally compact group but $G$ is not connected, and if $G_0$ is the connected component of the identity, then $G$ is the disjoint union of the cosets $gG_0$, and each coset $gG_0$ is homeomorphic to $G_0$, connected, and countable at infinity ($\sigma$-compact). This observation plays a crucial role in the proof of the uniqueness of the Haar measure (Theorem 6.21), because it guarantees that the use of Fubini’s theorem is legitimate.

The notion of uniform continuity can be generalized to functions defined on a group.

Definition 6.4. Given a topological group $G$ and a subset $S$ of $G$, for any normed vector space $F$, a function $f: G \to F$ is left uniformly continuous on $S$ if for any $\epsilon > 0$, there is an open subset $U$ of $G$ containing 1 such that

$$\|f(y) - f(x)\| < \epsilon \quad \text{for all } x, y \in S \text{ such that } y^{-1}x \in U.$$ 

The function $f: G \to F$ is right uniformly continuous on $S$ if for any $\epsilon > 0$, there is an open subset $U$ of $G$ containing 1 such that

$$\|f(y) - f(x)\| < \epsilon \quad \text{for all } x, y \in S \text{ such that } x^{-1}y \in U.$$ 

Observe that if $y^{-1}x \in U$, then we can write $y^{-1}x = z$ for some $z \in U$, so $y = z^{-1}x$, and $\|f(y) - f(x)\| = \|f(z^{-1}x) - f(x)\| < \epsilon$.

It is customary to introduce a left action $\lambda$ of $G$ on functions $f: G \to F$ defined on $G$ by

$$(\lambda_s(f))(x) = f(s^{-1}x) \quad \text{for all } x, s \in G.$$ 

Observe that

$$\lambda_{st}(f)(x) = f((st)^{-1}x) = f(t^{-1}s^{-1}x) = \lambda_t(f)(s^{-1}x) = \lambda_s(\lambda_t(f))(x),$$

where $\lambda_t$ is the left translation by $t$.


so
\[ \lambda_{st} = \lambda_s \circ \lambda_t, \]
which is the reason why we used \( s^{-1} \) instead of \( s \) in the definition of \( \lambda_s \).

Then \( \| f(z^{-1}x) - f(x) \| = \| \lambda_z(f)(x) - f(x) \| \), so the condition of the definition is equivalent to
\[ \| \lambda_z(f)(x) - f(x) \| < \epsilon \quad \text{for all } x \in S \text{ and all } z \in U. \]
Informally, the above condition can be written as
\[ \lim_{z \to 1} \sup_{x \in S} \| \lambda_z(f)(x) - f(x) \| = 0. \]

It is also customary to introduce a right action \( \rho \) of \( G \) on functions \( f : G \to F \) defined on \( G \) by
\[ (\rho_s(f))(x) = f(xs) \quad \text{for all } x, s \in G. \]
Observe that
\[ \rho_{st}(f)(x) = f(xst) = \rho_t(f)(xs) = \rho_s(\rho_t(f))(x), \]
so
\[ \rho_{st} = \rho_s \circ \rho_t. \]

Observe that if \( x^{-1}y \in U \), then we can write \( x^{-1}y = z \) for some \( z \in U \), so \( y = xz \), and
\[ \| f(y) - f(x) \| = \| f(xz) - f(x) \| = \| \rho_z(f)(x) - f(x) \| < \epsilon. \]

Thus the condition of the definition is equivalent to
\[ \| \rho_z(f)(x) - f(x) \| < \epsilon \quad \text{for all } x \in S \text{ and all } z \in U. \]
Informally, the above condition can be written as
\[ \lim_{z \to 1} \sup_{x \in S} \| \rho_z(f)(x) - f(x) \| = 0. \]

**Proposition 6.12.** Let \( G \) be a topological group. For any function \( f : S \to F \), where \( F \) is any normed vector space, if \( f \) is continuous with compact support \( K \), then \( f \) is left (resp. right) uniformly continuous on \( K \).

**Proof.** We prove that \( f \) is left uniformly continuous, the proof that \( f \) is right uniformly continuous being similar and left as an exercise. Since \( f \) is continuous, for every \( x \in K \), there is some open subset \( U_x \) with \( 1 \in U_x \) such that
\[ \| f(y) - f(x) \| < \frac{\epsilon}{2} \quad \text{for all } y \in xU_x. \]
We can find an open subset \( V_x \) containing \( 1 \) such that \( V_xV_x \subseteq U_x \). The open subsets of the form \( xV_x \) for \( x \in K \) form an open cover of \( K \), and since \( K \) is compact, there is a finite subcover \( \{ x_1V_{x_1}, \ldots, x_nV_{x_n} \} \) of \( K \) with \( x_1, \ldots, x_n \in K \). Let
\[ V = V_{x_1} \cap \cdots \cap V_{x_n}. \]
Consider $x, y \in K$ such that $y^{-1}x \in V$, that is, $x \in yV$. Then $y \in x_iV_i \subseteq x_iU_i$ for some $i$, and so
\[ x \in yV \in x_iV_i \subseteq x_iV_i x_i \subseteq x_i U_i, \]
which implies that
\[ \|f(x) - f(y)\| \leq \|f(x) - f(x_i)\| + \|f(x_i) - f(y)\| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon, \]
as desired. 

We end this section by combining the various properties of a topological group $G$ to characterize when $G/G_x$ is homeomorphic to $X$. In order to do so, we need two definitions.

**Definition 6.5.** Let $G$ be a topological group and let $X$ be a topological space. An action $\varphi: G \times X \to X$ is **continuous** (and $G$ acts **continuously on** $X$) if the map $\varphi$ is continuous.

If an action $\varphi: G \times X \to X$ is continuous, then each map $\varphi_g: X \to X$ is a homeomorphism of $X$ (recall that $\varphi_g(x) = g \cdot x$, for all $x \in X$).

Under some mild assumptions on $G$ and $X$, the quotient space $G/G_x$ is homeomorphic to $X$. For example, this happens if $X$ is a Baire space.

**Definition 6.6.** A **Baire space** $X$ is a topological space with the property that if $\{F_i\}_{i \geq 1}$ is any countable family of closed sets $F_i$ such that each $F_i$ has empty interior, then $\bigcup_{i \geq 1} F_i$ also has empty interior. By complementation, this is equivalent to the fact that for every countable family of open sets $U_i$ such that each $U_i$ is dense in $X$ (i.e., $\overline{U_i} = X$), then $\bigcap_{i \geq 1} U_i$ is also dense in $X$.

**Remark:** A subset $A \subseteq X$ is **rare** if its closure $\overline{A}$ has empty interior. A subset $Y \subseteq X$ is **meager** if it is a countable union of rare sets. Then, it is immediately verified that a space $X$ is a Baire space if and only if every nonempty open subset of $X$ is not meager.

The following theorem shows that there are plenty of Baire spaces:

**Theorem 6.13. (Baire)** (1) Every locally compact topological space is a Baire space.

(2) Every complete metric space is a Baire space.

A proof of Theorem 6.13 can be found in Bourbaki [8], Chapter IX, Section 5, Theorem 1.

**Theorem 6.14.** Let $G$ be a topological group which is locally compact and countable at infinity, $X$ a Hausdorff topological space which is a Baire space, and assume that $G$ acts transitively and continuously on $X$. Then, for any $x \in X$, the map $\varphi: G/G_x \to X$ is a homeomorphism.
Proof. We follow the proof given in Bourbaki [8], Chapter IX, Section 5, Proposition 6 (Essentially the same proof can be found in Mneimné and Testard [31], Chapter 2). First, observe that if a topological group acts continuously and transitively on a Hausdorff topological space, then for every \( x \in X \), the stabilizer \( G_x \) is a closed subgroup of \( G \). This is because, as the action is continuous, the projection \( \pi_x : G \rightarrow X : g \mapsto g \cdot x \) is continuous, and \( G_x = \pi^{-1}(\{x\}) \), with \( \{x\} \) closed. Therefore, by Proposition 6.5, the quotient space \( G/G_x \) is Hausdorff. As the map \( \pi_x : G \rightarrow X \) is continuous, the induced map \( \varphi_x : G/G_x \rightarrow X \) is continuous, and by Proposition 15.14, it is a bijection. Therefore, to prove that \( \varphi_x \) is a homeomorphism, it is enough to prove that \( \varphi_x \) is an open map. For this, it suffices to show that \( \pi_x \) is an open map. Given any open \( U \) in \( G \), we will prove that for any \( g \in U \), the element \( \pi_x(g) = g \cdot x \) is contained in the interior of \( U \cdot x \). However, observe that this is equivalent to proving that \( x \) belongs to the interior of \( (g^{-1} \cdot U) \cdot x \). Therefore, we are reduced to the following case: if \( U \) is any open subset of \( G \) containing \( 1 \), then \( x \) belongs to the interior of \( U \cdot x \).

Since \( G \) is locally compact, using Proposition 6.2, we can find a compact neighborhood of the form \( W = \overline{V} \), such that \( 1 \in W \), \( W = W^{-1} \) and \( W^2 \subseteq U \), where \( V \) is open with \( 1 \in V \subseteq U \). As \( G \) is countable at infinity, \( G = \bigcup_{i \geq 1} K_i \), where each \( K_i \) is compact. Since \( V \) is open, all the cosets \( gV \) are open, and as each \( K_i \) is covered by the \( gV \)'s, by compactness of \( K_i \), finitely many cosets \( gV \) cover each \( K_i \), and so

\[
G = \bigcup_{i \geq 1} g_i V = \bigcup_{i \geq 1} g_i W,
\]

for countably many \( g_i \in G \), where each \( g_i W \) is compact. As our action is transitive, we deduce that

\[
X = \bigcup_{i \geq 1} g_i W \cdot x,
\]

where each \( g_i W \cdot x \) is compact, since our action is continuous and the \( g_i W \) are compact. As \( X \) is Hausdorff, each \( g_i W \cdot x \) is closed, and as \( X \) is a Baire space expressed as a union of closed sets, one of the \( g_i W \cdot x \) must have nonempty interior; that is, there is some \( w \in W \), with \( g_i w \cdot x \) in the interior of \( g_i W \cdot x \), for some \( i \). But then, as the map \( y \mapsto g \cdot y \) is a homeomorphism for any given \( g \in G \) (where \( y \in X \)), we see that \( x \) is in the interior of

\[
w^{-1} g_i^{-1} \cdot (g_i W \cdot x) = w^{-1} W \cdot x \subseteq W^{-1} W \cdot x = W^2 \cdot x \subseteq U \cdot x,
\]

as desired. \( \square \)

By Theorem 6.13, we get the following important corollary:

**Theorem 6.15.** Let \( G \) be a topological group which is locally compact and countable at infinity, \( X \) a Hausdorff locally compact topological space, and assume that \( G \) acts transitively and continuously on \( X \). Then, for any \( x \in X \), the map \( \varphi_x : G/G_x \rightarrow X \) is a homeomorphism.

Readers who wish to learn more about topological groups may consult Sagle and Walde [37] and Chevalley [11] for an introductory account, and Bourbaki [7], Weil [44] and Pontryagin [34, 35], for a more comprehensive account (especially the last two references).
6.2 Existence of the Haar Measure

Let $G$ be a locally compact group. We are going to show there is a positive $\sigma$-regular locally finite Borel measure $\mu$ on $G$ such that $\mu(U) > 0$ for every nonempty open subset $U$, and such that $\mu$ is left-invariant, which means that

$$\mu(A) = \mu(L_x(A)) \quad \text{for all } x \in G \text{ and all } A \in B,$$

where $B$ is the $\sigma$-algebra of Borel sets on $G$.

Recall that for any $x \in G$, the maps $L_x : G \to G$ (left translation) and $R_x : G \to G$ (right translation) are defined by

$$L_x(z) = xz, \quad R_x(z) = zx, \quad \text{for all } x, z \in G.$$

It is obvious that $L_{xy} = L_x \circ L_y$ and $R_{xy} = R_y \circ R_x$, and that $L_x$ and $R_y$ commute for all $x, y \in G$.

It is customary to introduce a left action $\lambda$ of $G$ and a right action $\rho$ of $G$ on functions $f : G \to F$. We did this in the previous section, but for the sake of completeness, we repeat these definitions.

**Definition 6.7.** Let $G$ be a group, and let $F$ be any set. The left action $\lambda$ of $G$ on functions $f : G \to F$ is given by

$$\lambda_s(f)(x) = f(s^{-1}x) \quad \text{for all } x, s \in G,$$

and the right action $\rho$ of $G$ on functions $f : G \to F$ is given by

$$\rho_s(f)(x) = f(xs) \quad \text{for all } x, s \in G.$$

Observe that

$$\lambda_{st}(f)(x) = f(((st)^{-1}x) = f(t^{-1}s^{-1}x) = \lambda_t(f)(s^{-1}x) = \lambda_s(\lambda_t(f))(x),$$

so

$$\lambda_{st} = \lambda_s \circ \lambda_t,$$

which is the reason why we used $s^{-1}$ instead of $s$ in the definition of $\lambda_s$. Observe that

$$\rho_{st}(f)(x) = f(xst) = \rho_t(f)(xs) = \rho_s(\rho_t(f))(x),$$

so

$$\rho_{st} = \rho_s \circ \rho_t.$$

Given a subset $A$ of $G$, we usually write $sA$ for $L_s(A)$ and $As$ for $R_s(A)$.

We define a left action of $\lambda_s$ and a right action $\rho_s$ on measures and Radon functionals as follows.
Definition 6.8. Let $G$ be a locally compact topological group. The left action $\lambda$ of $G$ on a measure $\mu$ on $(G, \mathcal{B})$ is given by

$$(\lambda_s(\mu))(A) = \mu(s^{-1}A) \quad \text{for all } s \in G \text{ and all } A \in \mathcal{B},$$

and the right action $\rho$ of $G$ on a measure $\mu$ on $(G, \mathcal{B})$ is given by

$$(\rho_s(\mu))(A) = \mu(As) \quad \text{for all } s \in G \text{ and all } A \in \mathcal{B}.$$ 

The left action $\lambda$ of $G$ on a Radon functional $\Phi: \mathcal{K}_c(G) \to \mathbb{C}$ is given by

$$(\lambda_s(\Phi))(f) = \Phi(\lambda_s^{-1}(f)) \quad \text{for all } s \in G \text{ and all } f \in \mathcal{K}_c(G),$$

and the right action $\rho$ of $G$ on a Radon functional $\Phi: \mathcal{K}_c(G) \to \mathbb{C}$ is given by

$$(\rho_s(\Phi))(f) = \Phi(\rho_s^{-1}(f)) \quad \text{for all } s \in G \text{ and all } f \in \mathcal{K}_c(G).$$

If $m_{\Phi}$ is the Borel measure corresponding to a positive Radon functional $\Phi$, $m_{\lambda_s(\Phi)}$ is the Borel measure corresponding to $\lambda_s(\Phi)$, and $m_{\rho_s(\Phi)}$ is the Borel measure corresponding to $\rho_s(\Phi)$, given by Theorem 5.8, then we have

$$\int f(x)d(m_{\lambda_s(\Phi)})(x) = \int f(sx)dm_{\Phi}(x),$$

and

$$\int f(x)d(m_{\rho_s(\Phi)})(x) = \int f(xs^{-1})dm_{\Phi}(x)$$

for all $f \in \mathcal{K}_c(G)$ and all $s \in G$.

Definition 6.8 has been designed so that for every measure $\mu$ on $G$ we have

$$\lambda_{st}(\mu) = \lambda_s(\lambda_t(\mu)), \quad \rho_{st}(\mu) = \rho_s(\rho_t(\mu)) \quad \text{for all } s, t \in G.$$ 

For every Radon functional $\Phi$ and for every function $f \in \mathcal{K}_c(G)$, we have

$$(\lambda_{st}(\Phi))(f) = \Phi(\lambda_{(st)^{-1}}(f)) = \Phi(\lambda_{t^{-1}s^{-1}}(f))$$

$$= \Phi(\lambda_{t^{-1}}(\lambda_{s^{-1}}(f))) = (\lambda_t(\Phi))(\lambda_{s^{-1}}(f)) = (\lambda_s(\lambda_t(\Phi)))(f),$$

and a similar computation shows that

$$(\rho_{st}(\Phi))(f) = (\rho_s(\rho_t(\Phi)))(f).$$

Therefore, for every Radon functional $\Phi$, we have

$$\lambda_{st}(\Phi) = \lambda_s(\lambda_t(\Phi)), \quad \rho_{st}(\Phi) = \rho_s(\rho_t(\Phi)) \quad \text{for all } s, t \in G.$$
Definition 6.9. Let $G$ be a locally compact group. A left Haar measure $\mu$ is a $\sigma$-regular, locally finite, Borel measure on the $\sigma$-algebra $\mathcal{B}$ of Borel sets of $G$, such that $\mu(U) > 0$ for all nonempty open subsets $U \in \mathcal{B}$, and is left-invariant, which means that

$$\lambda_s(\mu) = \mu$$ for all $s \in G$.

The above condition means that

$$\mu(s^{-1}A) = \mu(A) \quad \text{for all } A \in \mathcal{B} \text{ and all } s \in G,$$

or equivalently,

$$\mu(sA) = \mu(A) \quad \text{for all } A \in \mathcal{B} \text{ and all } s \in G,$$

A right Haar measure $\mu$ is a Borel measure satisfying the same conditions as a left Haar measure, except that it is right invariant, which means that

$$\rho_s(\mu) = \mu$$ for all $A \in \mathcal{B}$ and all $s \in G$.

The above condition means that

$$\mu(As) = \mu(A) \quad \text{for all } A \in \mathcal{B} \text{ and all } s \in G.$$

In order to prove that a left (resp. right) Haar measure exists, we will use Theorem 5.8, which motivates the following definition.

Definition 6.10. Let $G$ be a locally compact group. A left Haar functional $\Phi$ is a positive non-zero Radon functional $\Phi : \mathcal{K}_C(G) \to \mathbb{C}$ which is left-invariant, which means that

$$\lambda_s(\Phi) = \Phi$$ for all $s \in G$.

A right Haar functional $\Phi$ is a positive non-zero Radon functional $\Phi : \mathcal{K}_C(G) \to \mathbb{C}$ which is right-invariant, which means that

$$\rho_s(\Phi) = \Phi$$ for all $s \in G$.

If $m_\Phi$ is the Borel measure associated with $\Phi$, then its left-invariance is equivalent to

$$\int f(x)dm_\Phi(x) = \int f(sx)dm_\Phi(x),$$

and its right-invariance is equivalent to

$$\int f(x)dm_\Phi(x) = \int f(xs^{-1})dm_\Phi(x),$$

for all $f \in \mathcal{K}_C(G)$ and all $s \in G$. 
**Definition 6.11.** Let \( G \) be any locally compact group, and \( F \) be any set. For any function \( f: G \to F \), define \( \hat{f}: G \to F \) by

\[
\hat{f}(s) = f(s^{-1}) \quad \text{for all } s \in G.
\]

For any Borel measure \( \mu \) on \((G, \mathcal{B})\), define \( \hat{\mu} \) by

\[
\hat{\mu}(A) = \mu(A^{-1}) \quad \text{for all } A \in \mathcal{B}.
\]

For any Radon functional \( \Phi: \mathcal{K}_C(G) \to \mathbb{C} \), define \( \hat{\Phi} : \mathcal{K}_C(G) \to \mathbb{C} \) by

\[
\hat{\Phi}(f) = \Phi(\hat{f}) \quad \text{for all } f \in \mathcal{K}_C(G).
\]

Observe that

\[
(\lambda_s(\mu))(A) = (\lambda_s(\mu))(A^{-1}) = \mu(s^{-1}A^{-1}) = \mu((As)^{-1}) = \hat{\mu}(As) = (\rho_s(\hat{\mu}))(A),
\]

so

\[
\lambda_s(\mu)^\sim = \rho_s(\hat{\mu}).
\]

Similarly,

\[
(\rho_s(\mu))(A) = (\rho_s(\mu))(A^{-1}) = \mu(A^{-1}s) = \mu((s^{-1}A)^{-1}) = \hat{\mu}(s^{-1}A) = (\lambda_s(\hat{\mu}))(A),
\]

so

\[
\rho_s(\mu)^\sim = \lambda_s(\hat{\mu}).
\]

For any function \( f: G \to F \), we have

\[
(\lambda_s(f))(x) = (\lambda_s(f))(x^{-1}) = f(s^{-1}x^{-1}) = f((xs)^{-1}) = \hat{f}(xs) = (\rho_s(\hat{f}))(x),
\]

and similarly

\[
(\rho_s(f))(x) = (\rho_s(f))(x^{-1}) = f(x^{-1}s) = f((s^{-1}x)^{-1}) = \hat{f}(s^{-1}x) = (\lambda_s(\hat{f}))(x).
\]

Therefore,

\[
\lambda_s(f)^\sim = \rho_s(\hat{f}), \quad \rho_s(f)^\sim = \lambda_s(\hat{f}) \quad \text{for all } s \in G \text{ and all } f: G \to F.
\]

Using the above equations, for every Radon functional \( \Phi: \mathcal{K}_C(G) \to \mathbb{C} \), we have

\[
(\lambda_s(\Phi))(f) = (\lambda_s(\Phi))(\hat{f}) = \Phi(\lambda_{s^{-1}}(\hat{f})) = \Phi(\rho_{s^{-1}}(f)^\sim) = \hat{\Phi}(\rho_{s^{-1}}(f)) = (\rho_s(\hat{\Phi}))(f).
\]

Similarly,

\[
(\rho_s(\Phi))(f) = (\rho_s(\Phi))(\hat{f}) = \Phi(\rho_{s^{-1}}(\hat{f})) = \Phi(\lambda_{s^{-1}}(f)^\sim) = \hat{\Phi}(\lambda_{s^{-1}}(f)) = (\lambda_s(\hat{\Phi}))(f).
\]

Therefore, we have

\[
\lambda_s(\Phi)^\sim = \rho_s(\hat{\Phi}), \quad \rho_s(\Phi)^\sim = \lambda_s(\hat{\Phi}) \quad \text{for all } s \in G.
\]
Proposition 6.16. Let $G$ be a locally compact group, and let $\mu$ be a $\sigma$-regular, locally finite, Borel measure on $G$. The following properties hold:

1. We have
   \[ \lambda_s(\mu) = \rho_s(\mu) = \lambda_s(\hat{\mu}) \quad \text{for all } s \in G. \]
   Consequently, $\mu$ is a left-invariant measure iff $\hat{\mu}$ is a right-invariant measure. For any Radon functional $\Phi: K_C(G) \to \mathbb{C}$, we have
   \[ \lambda_s(\Phi) = \rho_s(\hat{\Phi}), \quad \rho_s(\Phi) = \lambda_s(\hat{\Phi}) \quad \text{for all } s \in G. \]
   Consequently, $\Phi$ is left-invariant iff $\hat{\Phi}$ is right-invariant.

2. If the Haar measure $\mu$ is left-invariant then
   \[ \int \lambda_{s^{-1}}(f) \, d\mu = \int f \, d\mu \]
   for all $f \in L^1_\mu(G, \mathcal{B}, \mathbb{C})$ and all $s \in G$. If the Haar measure $\mu$ is right-invariant then
   \[ \int \rho_{s^{-1}}(f) \, d\mu = \int f \, d\mu \]
   for all $f \in L^1_\mu(G, \mathcal{B}, \mathbb{C})$ and all $s \in G$.

3. We have
   \[ \int \lambda_{s^{-1}}(f) \, d\mu = \int f d\lambda_s(\mu) \quad \text{for all } f \in L^1_\mu(G, \mathcal{B}, \mathbb{C}) \text{ and all } s \in G. \]
   If
   \[ \int \lambda_{s^{-1}}(f) \, d\mu = \int f \, d\mu \]
   for all $f \in K_C(G)$ and all $s \in G$, then $\mu$ is left-invariant. We have
   \[ \int \rho_{s^{-1}}(f) \, d\mu = \int f d\rho_s(\mu) \quad \text{for all } f \in L^1_\mu(G, \mathcal{B}, \mathbb{C}) \text{ and all } s \in G. \]
   If
   \[ \int \rho_{s^{-1}}(f) \, d\mu = \int f \, d\mu \]
   for all $f \in K_C(G)$ and all $s \in G$, then $\mu$ is right-invariant.

Proof. We already proved (1).

(2) Let $f$ be any $\mu$-step function
   \[ f = \sum_{k=1}^k y_k \chi_{A_k}, \]
where the $A_k$ are measurable Borel sets of finite measure. For all $s, x \in G$, we see that $(\lambda_{s^{-1}} f)(x) = f(sx) = y_k$ iff $sx \in A_k$ iff $x \in s^{-1}A_k$, which means that

$$\lambda_{s^{-1}} f = \sum_{k=1}^{k} y_k \chi_{s^{-1}A_k},$$

so

$$\int (\lambda_{s^{-1}} f) d\mu = \sum_{k=1}^{k} y_k \mu(s^{-1}A_k) = \sum_{k=1}^{k} y_k (\lambda_s(\mu))(A_k) = \int f d\lambda_s(\mu).$$

If $\mu$ is left-invariant, then $\lambda_s(\mu) = \mu$, so

$$\sum_{k=1}^{k} y_k (\lambda_s(\mu))(A_k) = \sum_{k=1}^{k} y_k \mu(A_k) = \int f d\mu,$$

and we deduce that

$$\int (\lambda_{s^{-1}} f) d\mu = \int f d\mu.$$

Every function $f \in L^1_{\mu}(G, \mathcal{B}, \mathbb{C})$ has some approximation sequence $(f_n)$ by $\mu$-step functions that converges to $f$ a.e. and in the $L^1$-norm. It follows that the sequence $(\lambda_{s^{-1}} f_n)$ converges a.e to $\lambda_{s^{-1}}(f)$. We check immediately that it is a Cauchy sequence because

$$\int (\lambda_{s^{-1}} f_n) d\mu = \int f_n d\mu$$

for all $n$, and it follows that

$$\int \lambda_{s^{-1}}(f) d\mu = \int f d\mu.$$

The case where $\mu$ is right-invariant is analogous.

(3) The proof in (2) actually shows that

$$\int \lambda_{s^{-1}}(f) d\mu = \int f d\lambda_s(\mu).$$

If

$$\int \lambda_{s^{-1}}(f) d\mu = \int f d\mu$$

then

$$\int f d\lambda_s(\mu) = \int f d\mu$$

for all $f \in K_{\mathbb{C}}(G)$ and all $s \in G$, and by the uniqueness of the Borel measure corresponding to the Radon functional $f \mapsto \int f d\mu$ from Theorem 5.8, we see that $\lambda_s(\mu) = \mu$ for all $s \in G$, which means that $\mu$ is left-invariant.
The condition
\[ \int \lambda_{s^{-1}}(f) \, d\mu = \int f \, d\mu \quad \text{for all } s \in G \]
is also written as
\[ \int f(sx) \, d\mu(x) = \int f(x) \, d\mu(x) \quad \text{for all } s \in G. \]

The condition
\[ \int \rho_{s^{-1}}(f) \, d\mu = \int f \, d\mu \quad \text{for all } s \in G \]
is also written as
\[ \int f(xs^{-1}) \, d\mu(x) = \int f(x) \, d\mu(x) \quad \text{for all } s \in G. \]

Since \( G \) is a group and \( s \) is any arbitrary element of \( G \), the above condition is also equivalent to
\[ \int f(xs) \, d\mu(x) = \int f(x) \, d\mu(x) \quad \text{for all } s \in G. \]

We are now going to sketch the proof that a left-invariant Haar measure exists on any locally compact group. All proofs we are aware of (Weil [44], Halmos [22], Bourbaki [4], Dieudonné [12], Lang [24], Folland [18]) make use of Haar’s original clever idea (1933). Except for Halmos who constructs directly a measure (as Haar did), all the other proofs are essentially André Weil’s proof (which constructs a Haar functional) from his famous little book [44] first published in 1940.

**Theorem 6.17. (Haar)** Every locally compact group \( G \) possesses a left-invariant Haar measure.

**Proof sketch.** Folland [18] (Chapter 2, Section 2.2) is kind enough to provide the intuition behind the construction. In this method, a measure is not constructed directly. Instead, a Haar functional is constructed. Then Theorem 5.8 is used to obtain a left-invariant Borel measure which is a Haar measure.

Suppose we have positive function \( \varphi \in \mathcal{K}_{\mathbb{R}}(G) \) bounded by 1, equal to 1 on a small open set \( U \), and whose support is a compact subset slightly larger than \( U \). If \( f \in \mathcal{K}_{\mathbb{R}}(G) \) is any other function slowly varying so that it is essentially constant on the left translates of \( U \), then \( f \) can be approximated by a linear combination \( f \approx \sum c_j \lambda_{g_j}(\varphi) \). If \( \mu \) were a left-invariant Haar measure, then we would have
\[ \int f \, d\mu \approx \left( \sum_j c_j \right) \int \varphi \, d\mu. \]

This approximation gets better and better as the support of \( \varphi \) shrinks to a point, and if we introduce a normalization to cancel out the factor \( \int \varphi \, d\mu \), then we obtain \( \int f \, d\mu \) as the limit of the sums \( \sum_j c_j \). If \( G = \mathbb{R} \), and if \( \varphi \) is the characteristic function of a small interval, this
is reminiscent of approximation of \( \int f \, d\mu \) by Riemann sums. The issue is to make this idea precise and formal.

The first step is to pick some positive nonzero function \( \varphi \in \mathcal{K}_\mathbb{R}(G) \). Then we claim that for every function \( f \in \mathcal{K}_\mathbb{R}(G) \), there exists a finite set \( \{g_1, \ldots, g_n\} \) of elements of \( G \) and a finite sequence \((c_1, \ldots, c_n)\) of reals \( c_j \in \mathbb{R} \), such that

\[
  f \leq \sum_{j=1}^n c_j \lambda_{g_j}(\varphi).
\]

This is because \( f \) has compact support, so its support can be covered by finitely many translates of the open subset \( U \) given by

\[
  U = \left\{ g \in G \mid \varphi(g) > \frac{1}{2} \|\varphi\|_\infty \right\},
\]

and if we pick \( c_j = \|f\|_\infty / a \) where \( a = (1/2) \|\varphi\|_\infty \), then \( f \leq \sum_{j=1}^n c_j \lambda_{g_j}(\varphi) \).

Define \( (f : \varphi) \) as the greatest lower bound of the sums \( \sum_{j=1}^n c_j \), over all set \( \{g_1, \ldots, g_n\} \) of elements of \( G \) and all finite sequence \((c_1, \ldots, c_n)\) of reals \( c_j \in \mathbb{R} \) such that

\[
  f \leq \sum_{j=1}^n c_j \lambda_{g_j}(\varphi).
\]

Then it is not hard to show that the quantity \( (f : \varphi) \) has the following properties:

- \( (f : \varphi) = (\gamma_s(f) : \varphi) \) for all \( s \in G \)
- \( (f_1 + f_2 : \varphi) \leq (f_1 : \varphi) + (f_2 : \varphi) \) for all \( f_1, f_2 \in \mathcal{K}_\mathbb{R}(G) \)
- \( (cf : \varphi) = c(f : \varphi) \) for all \( c \geq 0 \)
- \( (f_1 : \varphi) \leq (f_2 : \varphi) \) whenever \( f_1 \leq f_2 \)
- \( (f : \varphi) \geq \|f\|_\infty / \|\varphi\|_\infty \)
- \( (f : \varphi) \leq (f : \psi)(\psi : \varphi) \) for all positive \( \psi \in \mathcal{K}_\mathbb{R}(G) \)

\[
  0 < \frac{1}{(f_0 : f)} \leq \frac{(f : \varphi)}{(f_0 : \varphi)} \leq (f : f_0) \leq (f : f_0) \quad \text{for all positive } f, f_0 \in \mathcal{K}_\mathbb{R}(G).
\]

We now make a normalization by fixing some positive nonzero \( f_0 \in \mathcal{K}_\mathbb{R}(G) \), and defining

\[
  I_\varphi(f) = \frac{(f : \varphi)}{(f_0 : \varphi)}
\]

for every positive function \( f \in \mathcal{K}_\mathbb{R}(G) \). The above properties show that \( I_\varphi \) is a functional which is left-invariant, subaddtive, homogeneous of degree 1, and monotone. It also satisfies the following property:

\[
  \frac{1}{(f_0 : f)} \leq I_\varphi(f) \leq (f : f_0).
\]

(*)
If \( I_\varphi \) were additive rather than subadditive, it would be the restriction to the positive functions in \( K_\mathbb{R}(G) \) of a positive linear functional on \( K_\mathbb{R}(G) \), and we would be done. To make a linear functional, we need to shrink to the domain of \( \varphi \), and this is the part of the argument which is the most subtle.

The following technical proposition whose proof is given in Folland [18] (Chapter 2, Lemma 2.18) is needed.

**Proposition 6.18.** If \( f_1 \) and \( f_2 \) are any two positive functions in \( K_\mathbb{R}(G) \) and if \( \epsilon > 0 \), then there is an open subset \( V \) containing 1 such that \( I_\varphi(f_1) + I_\varphi(f_2) \leq I_\varphi(f_1 + f_2) + \epsilon \), whenever \( \text{supp}(\varphi) \subseteq V \).

To shrink the domain of \( \varphi \) we use a compactness argument. For every positive function \( f \in K_\mathbb{R}(G) \), let \( X_f \) be the interval

\[
X_f = \left[ \frac{1}{(f_0 : f)}, (f : f_0) \right].
\]

Let \( X = \prod_f X_f \). Since each \( X_f \) is compact, by Tychonoff’s theorem \( X \) is also compact. \( \mathbb{W} \) can think of \( X \) as the set of functions mapping a positive function \( f \in K_\mathbb{R}(G) \) to some number in the interval \( X_f \). By (*), we have \( I_\varphi(f) \in X \) for all positive nonzero \( \varphi \in K_\mathbb{R}(G) \). For every open subset \( V \) containing 1, let \( K(V) \) be the closure in \( X \) of the subset \( \{ I_\varphi \mid \text{supp}(\varphi) \subseteq V \} \). The family of subset \( K(V) \) has the finite intersection property, since \( K(\bigcap_{j=1}^n V_j) \subseteq \bigcap_{j=1}^n K(V_j) \). Since \( X \) is compact, there is some \( I \in X \) which lies in every \( K(V) \). This means that every neighborhood of \( I \) in \( X \) contains some \( I_\varphi \) with \( \text{supp}(\varphi) \) arbitrarily small. In other words, for any open subset \( V \) containing 1, and for any positive functions \( f_1, \ldots, f_n \in K_\mathbb{R}(G) \), there exist some positive nonzero \( \varphi \in K_\mathbb{R}(G) \) with \( \text{supp}(\varphi) \subseteq V \) such that \( |I(f_j) - I_\varphi(f_j)| < \epsilon \) for all \( j \). By the properties of \( I_\varphi \) listed above and by Proposition 6.18, we conclude that \( I \) commutes with left translation, addition, and multiplication by positive scalars.

We can extend \( I \) to arbitrary functions \( f \in K_\mathbb{R}(G) \) as follows. We can write \( f = f_1 - f_2 \) with \( f_1, f_2 \) positive functions in \( K_\mathbb{R}(G) \), and we let \( I(f) = I(f_1) - I(f_2) \). If we also have \( f = f'_1 - f'_2 \) with \( f'_1, f'_2 \) positive functions in \( K_\mathbb{R}(G) \), then

\[
f_1 + f'_2 = f_2 + f'_1,
\]

so by linearity of \( I \) on positive functions we get

\[
I(f_1) + (f'_2) = I(f_2) + I(f'_1),
\]

thus

\[
I(f_1) - I(f_2) = I(f'_1) - I(f'_2),
\]

which means that \( I(f) \) is well defined. The functional \( I \) is a left Haar functional, and we are done. \( \square \)
Remark: The proof in Bourbaki [4] uses an argument involving an ultrafilter instead of Tychonoff’s theorem, but otherwise it is identical. Dieudonné [12] assumes that the locally compact group $G$ is separable and metrizable. This allows him to avoid using Tychonoff’s theorem, but does not make the proof simpler.

Let $\mu$ be the left Haar measure associated with the left Haar functional $I$ given by Theorem 6.17. Here is an immediate consequence of Theorem 6.17.

**Proposition 6.19.** If $\mu$ is left Haar measure on $G$, then for every nonempty open subset $U$, we have $\mu(U) > 0$. For every positive nonzero function $f \in K_R(G)$, we have $\int fd\mu > 0$.

**Proof.** Assume $U$ is a nonempty open set with $\mu(U) = 0$. Then since $\mu$ is left-invariant $\mu(gU) = 0$ for all $g \in G$, and since any compact subset $K$ can be covered by finitely many translates of $U$, we have $\mu(K) = 0$. Since $\mu$ is a regular Borel measure, it is inner regular, that is,

$$\mu(E) = \sup \{\mu(K) \mid K \subseteq E, K \text{ compact}\}$$

for any $E \in \mathcal{B}$, so $\mu(E) = 0$ for all $E \in \mathcal{B}$, contradicting the fact that $\mu$ is not the zero measure.

For any positive nonzero function $f \in K_R(G)$, let $U = \{g \in G \mid f(g) > \frac{1}{2} \|f\|_\infty\}$. Then $\int fd\mu > \frac{1}{2} \|f\|_\infty \mu(U) > 0$. \qed

We now turn to the uniqueness of the Haar measure.

### 6.3 Uniqueness of the Haar Measure

Any two left Haar measures on a locally compact group are proportional up to a positive factor. All the proofs we are aware of use tricks involving a double integration and Fubini’s theorem. These proofs are attributed to von Neumann. In our opinion, the proof using the least devious trick is that of Dieudonné [12] (Chapter XIV, Section 1), also used in Bourbaki [4] in a slightly more concise form. Since Dieudonné uses a theory of integration based on Radon functionals rather than on measure theory, some minor adaptations need to be made; specifically, Proposition 5.14 is needed instead of Proposition 13.15.3 in Dieudonné [12]. The first step is the following crucial result.

**Proposition 6.20.** Given a left Haar functional $\Phi$ and a right Haar functional $\Psi$ on a locally compact group $G$, if $\nu$ is the corresponding right Haar measure, for any function $f \in K_R(G)$, if $\Phi(f) \neq 0$, then the function $D_f$ given by

$$D_f(s) = \Phi(f)^{-1} \int f(t^{-1}s) d\nu(t)$$

for all $s \in G$ is continuous.
6.3. **UNIQUENESS OF THE HAAR MEASURE**

*Proof.* It suffices to show that the function

\[ s \mapsto \int f(t^{-1}s) \, d\nu(t) \]

is continuous. Let \( K \) be the compact subset of \( G \) which is the support of \( f \). Pick any \( s_0 \in G \), and let \( V_0 \) by any compact neighborhood of \( s_0 \). For every \( \epsilon > 0 \), we have to find an open subset \( V \) containing \( s_0 \) such that

\[ \left| \int (f(t^{-1}s) - f(t^{-1}s_0)) \, d\nu \right| < \epsilon \]

for all \( s \in V \). Let \( L = KV_0^{-1} \), then

\[ \int (f(t^{-1}s) - f(t^{-1}s_0)) \, d\nu = \int_L (f(t^{-1}s) - f(t^{-1}s_0)) \, d\nu. \]

By Proposition 6.12, the function \( f \) is right uniformly continuous, so there is some open subset \( W \) containing 1 such that

\[ |f(t^{-1}s) - f(t^{-1}s_0)| < \frac{\epsilon}{\nu(L)} \]

for all \((t^{-1}s_0)^{-1}t^{-1}s \in W\), that is, \( s_0^{-1}s \in W \), namely \( s \in Ws_0 \), and all \( t \in G \). If we take \( V = V_0 \cap W \), then

\[ \int_L (f(t^{-1}s) - f(t^{-1}s_0)) \, d\nu < \int_L |f(t^{-1}s) - f(t^{-1}s_0)| \, d\nu < \epsilon, \]

as desired. \( \square \)

We are now ready to prove our uniqueness result.

**Theorem 6.21.** (Haar) If \( \mu \) and \( \nu \) are any two left-invariant Haar measures on a locally compact group \( G \), then there is some \( c > 0 \) such that \( \mu = c\nu \).

*Proof.* Since a Haar functional \( \Psi \) is right-invariant iff \( \Psi \) is left-invariant, It suffices to prove that if \( \Phi \) is a left-invariant Haar functional and if \( \Psi \) is a right-invariant Haar functional, then there is some \( c > 0 \) such that \( \Phi = c\Psi \). Let \( \mu \) be the left Haar measure associated with \( \Phi \) and let \( \nu \) be right Haar measure associated with \( \Psi \).

Let \( f \in \mathcal{K}_R(G) \) be any function such that \( \Phi(f) \neq 0 \) and let \( g \in \mathcal{K}_R(G) \) be any other function. The function from \( G \times G \) to \( \mathbb{R} \) given by \((s,t) \mapsto f(s)g(ts)\) is continuous and has compact support. Recall that \( D_f \) is given by

\[ D_f(s) = \Phi(f)^{-1} \int f(t^{-1}s) \, d\nu(t) \]
We evaluate $\Phi(f)\Psi(g)$ using Fubini’s theorem.

\[
\Phi(f)\Psi(g) = \left(\int f(s)d\mu(s)\right)\left(\int g(t)d\nu(t)\right)
\]
\[
= \int f(s)\left(\int g(t)d\nu(t)\right)d\mu(s)
\]
\[
= \int \left(\int f(s)g(t)d\nu(t)\right)d\mu(s)
\]
\[
= \int \left(\int f(s)g(ts)d\nu(t)\right)d\mu(s) \quad \text{by right-invariance of } \nu
\]
\[
= \int \left(\int f(s)g(ts)d\mu(s)\right)d\nu(t) \quad \text{by Fubini}
\]
\[
= \int \left(\int f(t^{-1}s)g(s)d\mu(s)\right)d\nu(t) \quad \text{by left-invariance of } \mu
\]
\[
= \int \left(\int f(t^{-1}s)g(s)d\nu(t)\right)d\mu(s) \quad \text{by Fubini}
\]
\[
= \int g(s)\Phi(f)D_f(s)d\mu(s) \quad \text{by definition of } D_f
\]
\[
= \Phi(f)\Phi(D_f \cdot g).
\]

Since $\Phi(f) \neq 0$, we deduce that

\[
\Psi(g) = \Phi(D_f \cdot g).
\]

The above equation shows that $D_f$ is independent of $f$ because if $f'$ is another function $f' \in \mathcal{K}_\mathbb{R}(G)$ such that $\Phi(f') \neq 0$, then

\[
\Phi(D_f \cdot g) = \int D_f(s)g(s)d\mu(s) = \int D_f'(s)g(s)d\mu(s) = \Phi(D_f' \cdot g) \quad \text{for all } g \in \mathcal{K}_\mathbb{R}(G).
\]

By Proposition 5.14, we deduce that $D_f$ and $D_{f'}$ are equal a.e. (The version of Proposition 5.14 is stated for complex-valued functions, but it also holds for real-valued functions). However, $D_f$ and $D_{f'}$ are continuous and the subset $N$ where they differ is open and a null set, thus empty. Therefore $D_f = D_{f'} = D$, and by definition of $D$ we have

\[
\Phi(f) = D(1)\Psi(f) \quad \text{for all } f \in \mathcal{K}_\mathbb{R}(G) \text{ with } \Phi(f) \neq 0.
\]

Now $\Phi$ and $\Psi$ are two linear functionals that agree in the complement of the hyperplane $H$ in $\mathcal{K}_\mathbb{R}(G)$ of equation $\Phi(f) = 0$, so they agree everywhere. To see this, pick a basis of $\mathcal{K}_\mathbb{R}(G)$ consisting of a basis $(h_j)_{j \in J}$ of $H$ and a function $v$ not in $H$. We claim that the family consisting of $(h_j + v)_{j \in J}$ and $v$ is a basis of $\mathcal{K}_\mathbb{R}(G)$. This family obviously spans $\mathcal{K}_\mathbb{R}(G)$ (since
every \( h_j \) is obtained as \( h_j + v - v \), and it is linearly independent because if we have a finite linear combination

\[
\sum_{i \in I} \lambda_i (h_i + v) + \mu v = 0.
\]

for any finite subset \( I \) of \( J \), then

\[
\sum_{i \in I} \lambda_i h_i + \left( \mu + \sum_{i=1}^{\lfloor |I| \rfloor} \lambda_i \right) v = 0,
\]

and by linear independence, \( \lambda_i = 0 \) for all \( i \in I \) and \( \mu + \sum_{i=1}^{\lfloor |I| \rfloor} \lambda_i = 0 \), which implies \( \mu = 0 \), and since this holds for any finite subset \( I \) of \( J \), the family consisting of \( (h_j + v)_{j \in J} \) and \( v \) is linearly independent. Since \( \Phi \) and \( \Psi \) are linear and they agree on a basis, they must be identical.

Since \( \Psi \neq 0 \), we must have \( D(1) \neq 0 \), thus, \( \Phi = D(1)\Psi \). Since \( \Phi \) and \( \Psi \) are positive functionals, we must have \( D(1) > 0 \).

As we observed earlier, since a locally compact group is the disjoint union of \( \sigma \)-compact cosets, it is legitimate to use Fubini’s theorem.

In the next section, we explore the relationship between a left Haar measure \( \mu \) and the left Haar measure \( \rho_s(\mu) \).

### 6.4 The Modular Function

### 6.5 Examples of Haar Measures

### 6.6 Haar Measure on a Quotient Group

### 6.7 Convolution
Chapter 7

The Fourier Transform on \( \mathbb{R}^n \)
Chapter 8

Normed Algebras and Spectral Theory

8.1 Normed Algebras
8.2 Spectrum in a Normed Algebra
8.3 Characters and Spectrum in a Normed Algebra
8.4 Banach Algebras with Involution
8.5 Representations of Banach Algebras with Involution
8.6 Positive Linear Forms and Positive Hilbert Forms
8.7 Traces, Bitraces, Hilbert Algebras
8.8 Complete Hilbert Algebras
8.9 The Plancherel–Godement Theorem
Chapter 9

Unitary Representations of Locally Compact Groups

9.1 Group Representations

9.2 Representations of $G$ and Representations of $L^1(G)$

9.3 Functions of Positive Type
Chapter 10
Locally Compact Abelian Groups

10.1 Characters
10.2 The Dual Group
10.3 The Fourier Transform
10.4 Pontrjagin Duality
10.5 Spectral Synthesis
Chapter 11

Representations of Compact Groups

11.1 The Peter–Weyl Theorem

11.2 The Fourier Transform
Chapter 12

Gelfand Pairs

12.1 Characters

12.2 The Fourier Transform
Chapter 13

Topology

13.1 Metric Spaces and Normed Vector Spaces

This chapter contains a review of basic topological concepts. First metric spaces are defined. Next normed vector spaces are defined. Closed and open sets are defined, and their basic properties are stated. The general concept of a topological space is defined. The closure and the interior of a subset are defined. The subspace topology and the product topology are defined. Continuous maps and homeomorphisms are defined. Limits of sequences are defined. Continuous linear maps and multilinear maps are defined and studied briefly.

Most spaces considered in this book have a topological structure given by a metric or a norm, and we first review these notions. We begin with metric spaces. Recall that $\mathbb{R}_+ = \{x \in \mathbb{R} \mid x \geq 0\}$.

Definition 13.1. A metric space is a set $E$ together with a function $d: E \times E \to \mathbb{R}_+$, called a metric, or distance, assigning a nonnegative real number $d(x, y)$ to any two points $x, y \in E$, and satisfying the following conditions for all $x, y, z \in E$:

(D1) $d(x, y) = d(y, x)$. (symmetry)

(D2) $d(x, y) \geq 0$, and $d(x, y) = 0$ iff $x = y$. (positivity)

(D3) $d(x, z) \leq d(x, y) + d(y, z)$. (triangle inequality)

Geometrically, Condition (D3) expresses the fact that in a triangle with vertices $x, y, z$, the length of any side is bounded by the sum of the lengths of the other two sides. From (D3), we immediately get

$$|d(x, y) - d(y, z)| \leq d(x, z).$$

Let us give some examples of metric spaces. Recall that the absolute value $|x|$ of a real number $x \in \mathbb{R}$ is defined such that $|x| = x$ if $x \geq 0$, $|x| = -x$ if $x < 0$, and for a complex number $x = a + ib$, by $|x| = \sqrt{a^2 + b^2}$. 

163
Example 13.1.

1. Let $E = \mathbb{R}$, and $d(x, y) = |x - y|$, the absolute value of $x - y$. This is the so-called natural metric on $\mathbb{R}$.

2. Let $E = \mathbb{R}^n$ (or $E = \mathbb{C}^n$). We have the Euclidean metric
   \[ d_2(x, y) = (|x_1 - y_1|^2 + \cdots + |x_n - y_n|^2)^{\frac{1}{2}}, \]
   the distance between the points $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$.

3. For every set $E$, we can define the discrete metric, defined such that $d(x, y) = 1$ iff $x \neq y$, and $d(x, x) = 0$.

4. For any $a, b \in \mathbb{R}$ such that $a < b$, we define the following sets:
   \[ [a, b] = \{ x \in \mathbb{R} \mid a \leq x \leq b \}, \quad (\text{closed interval}) \]
   \[ (a, b) = \{ x \in \mathbb{R} \mid a < x < b \}, \quad (\text{open interval}) \]
   \[ [a, b) = \{ x \in \mathbb{R} \mid a \leq x < b \}, \quad (\text{interval closed on the left, open on the right}) \]
   \[ (a, b] = \{ x \in \mathbb{R} \mid a < x \leq b \}, \quad (\text{interval open on the left, closed on the right}) \]

   Let $E = [a, b]$, and $d(x, y) = |x - y|$. Then, $([a, b], d)$ is a metric space.

We will need to define the notion of proximity in order to define convergence of limits and continuity of functions. For this we introduce some standard “small neighborhoods.”

Definition 13.2. Given a metric space $E$ with metric $d$, for every $a \in E$, for every $\rho \in \mathbb{R}$, with $\rho > 0$, the set
   \[ B(a, \rho) = \{ x \in E \mid d(a, x) \leq \rho \} \]
   is called the \textit{closed ball of center $a$ and radius $\rho$}, the set
   \[ B_0(a, \rho) = \{ x \in E \mid d(a, x) < \rho \} \]
   is called the \textit{open ball of center $a$ and radius $\rho$}, and the set
   \[ S(a, \rho) = \{ x \in E \mid d(a, x) = \rho \} \]
   is called the \textit{sphere of center $a$ and radius $\rho$}. It should be noted that $\rho$ is finite (i.e., not $+\infty$). A subset $X$ of a metric space $E$ is \textit{bounded} if there is a closed ball $B(a, \rho)$ such that $X \subseteq B(a, \rho)$.

Clearly, $B(a, \rho) = B_0(a, \rho) \cup S(a, \rho)$.
Example 13.2.

1. In $E = \mathbb{R}$ with the distance $|x - y|$, an open ball of center $a$ and radius $\rho$ is the open interval $(a - \rho, a + \rho)$.

2. In $E = \mathbb{R}^2$ with the Euclidean metric, an open ball of center $a$ and radius $\rho$ is the set of points inside the disk of center $a$ and radius $\rho$, excluding the boundary points on the circle.

3. In $E = \mathbb{R}^3$ with the Euclidean metric, an open ball of center $a$ and radius $\rho$ is the set of points inside the sphere of center $a$ and radius $\rho$, excluding the boundary points on the sphere.

One should be aware that intuition can be misleading in forming a geometric image of a closed (or open) ball. For example, if $d$ is the discrete metric, a closed ball of center $a$ and radius $\rho < 1$ consists only of its center $a$, and a closed ball of center $a$ and radius $\rho \geq 1$ consists of the entire space!

If $E = [a, b]$, and $d(x, y) = |x - y|$, as in Example 13.1, an open ball $B_0(a, \rho)$, with $\rho < b - a$, is in fact the interval $[a, a + \rho)$, which is closed on the left.

We now consider a very important special case of metric spaces, normed vector spaces. Normed vector spaces have already been defined in Chapter 14 (Definition 14.1) but for the reader’s convenience we repeat the definition.

**Definition 13.3.** Let $E$ be a vector space over a field $K$, where $K$ is either the field $\mathbb{R}$ of reals, or the field $\mathbb{C}$ of complex numbers. A norm on $E$ is a function $\|\| : E \to \mathbb{R}_+$ assigning a nonnegative real number $\|u\|$ to any vector $u \in E$, and satisfying the following conditions for all $x, y, z \in E$:

(N1) $\|x\| \geq 0$, and $\|x\| = 0$ iff $x = 0$. \hspace{1cm} \text{(positivity)}

(N2) $\|\lambda x\| = |\lambda| \|x\|$. \hspace{1cm} \text{(scaling)}

(N3) $\|x + y\| \leq \|x\| + \|y\|$. \hspace{1cm} \text{(triangle inequality)}

A vector space $E$ together with a norm $\|\|$ is called a normed vector space. A function $\|\| : E \to \mathbb{R}_+$ satifying only properties (N2) and (N3) is called a semi-norm.

From (N3), we easily get

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

Given a normed vector space $E$, if we define $d$ such that

$$d(x, y) = \|x - y\|,$$
it is easily seen that $d$ is a metric. Thus, every normed vector space is immediately a metric space. Note that the metric associated with a norm is invariant under translation, that is,

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

For this reason, we can restrict ourselves to open or closed balls of center 0.

If $\|\cdot\| : E \to \mathbb{R}_+$ is a semi-norm, then $\|x\| = 0$ does not necessarily imply that $x = 0$. However, by setting $\lambda = 0$ and $x = 0$ in (N2), we see that $\|0\| = 0$. If we let $\mathcal{N} = \{x \in E \mid \|x\| = 0\}$, then $\mathcal{N}$ is a subspace of $E$. Indeed, $0 \in \mathcal{N}$, and if $\|x\| = \|y\| = 0$, then by (N2) and (N3) we have

$$\|\lambda x + \mu y\| \leq \|\lambda x\| + \|\mu y\| = |\lambda| \|x\| + |\mu| \|y\| = 0 + 0 = 0,$$

so $\lambda x + \mu y \in \mathcal{N}$. We can form the quotient space $E/\mathcal{N}$, and then it is easy to see that the semi-norm $\|\cdot\|$ induces a norm on $X/\mathcal{N}$.

Natural examples of semi-norms arise in integration theory; see Chapter 4.

Examples of normed vector spaces were given in Example 14.1. We repeat the most important examples.

**Example 13.3.** Let $E = \mathbb{R}^n$ (or $E = \mathbb{C}^n$). There are three standard norms. For every $(x_1, \ldots, x_n) \in E$, we have the norm $\|x\|_1$, defined such that,

$$\|x\|_1 = |x_1| + \cdots + |x_n|,$$

we have the *Euclidean norm* $\|x\|_2$, defined such that,

$$\|x\|_2 = (|x_1|^2 + \cdots + |x_n|^2)^{\frac{1}{2}},$$

and the *sup-norm* $\|x\|_\infty$, defined such that,

$$\|x\|_\infty = \max\{|x_i| \mid 1 \leq i \leq n\}.$$

More generally, we define the $\ell_p$-norm (for $p \geq 1$) by

$$\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}.$$
13.1. METRIC SPACES AND NORMED VECTOR SPACES

Figure 13.1: Figure (a) shows the diamond shaped closed ball associated with $\| \cdot \|_1$. Figure (b) shows the closed unit disk associated with $\| \cdot \|_2$, while Figure (c) illustrates the closed unit ball associated with $\| \cdot \|_\infty$.

Figure 13.2: The relationship between the closed unit balls centered at $(0,0)$. 
In a normed vector space we define a closed ball or an open ball of radius $\rho$ as a closed ball or an open ball of center 0. We may use the notation $B(\rho)$ and $B_0(\rho)$.

We will now define the crucial notions of open sets and closed sets, and of a topological space.

**Definition 13.4.** Let $(E,d)$ be a metric space. A subset $U \subseteq E$ is an open set in $E$ if either $U = \emptyset$, or for every $a \in U$, there is some open ball $B_0(a,\rho)$ such that, $B_0(a,\rho) \subseteq U$.\(^1\) A subset $F \subseteq E$ is a closed set in $E$ if its complement $E - F$ is open in $E$. See Figure 13.5.

The set $E$ itself is open, since for every $a \in E$, every open ball of center $a$ is contained in $E$. In $E = \mathbb{R}^n$, given $n$ intervals $[a_i,b_i]$, with $a_i < b_i$, it is easy to show that the open $n$-cube

$$\{(x_1,\ldots,x_n) \in E \mid a_i < x_i < b_i, \ 1 \leq i \leq n\}$$

\(^1\)Recall that $\rho > 0$.  

---

*Figure 13.4: The relationship between the closed unit balls centered at $(0,0,0)$.  

Figure 13.5: An open set $U$ in $E = \mathbb{R}^2$ under the standard Euclidean metric. Any point in the peach set $U$ is surrounded by a small raspberry open set which lies within $U$.  

The set $E$ itself is open, since for every $a \in E$, every open ball of center $a$ is contained in $E$. In $E = \mathbb{R}^n$, given $n$ intervals $[a_i,b_i]$, with $a_i < b_i$, it is easy to show that the open $n$-cube*
is an open set. In fact, it is possible to find a metric for which such open $n$-cubes are open balls! Similarly, we can define the closed $n$-cube

$$\{(x_1, \ldots, x_n) \in E \mid a_i \leq x_i \leq b_i, 1 \leq i \leq n\},$$

which is a closed set.

The open sets satisfy some important properties that lead to the definition of a topological space.

**Proposition 13.1.** Given a metric space $E$ with metric $d$, the family $\mathcal{O}$ of all open sets defined in Definition 13.4 satisfies the following properties:

(01) For every finite family $(U_i)_{1 \leq i \leq n}$ of sets $U_i \in \mathcal{O}$, we have $U_1 \cap \cdots \cap U_n \in \mathcal{O}$, i.e., $\mathcal{O}$ is closed under finite intersections.

(02) For every arbitrary family $(U_i)_{i \in I}$ of sets $U_i \in \mathcal{O}$, we have $\bigcup_{i \in I} U_i \in \mathcal{O}$, i.e., $\mathcal{O}$ is closed under arbitrary unions.

(03) $\emptyset \in \mathcal{O}$, and $E \in \mathcal{O}$, i.e., $\emptyset$ and $E$ belong to $\mathcal{O}$.

Furthermore, for any two distinct points $a \neq b$ in $E$, there exist two open sets $U_a$ and $U_b$ such that, $a \in U_a$, $b \in U_b$, and $U_a \cap U_b = \emptyset$.

**Proof.** It is straightforward. For the last point, letting $\rho = d(a, b)/3$ (in fact $\rho = d(a, b)/2$ works too), we can pick $U_a = B_0(a, \rho)$ and $U_b = B_0(b, \rho)$. By the triangle inequality, we must have $U_a \cap U_b = \emptyset$. \hfill $\square$

The above proposition leads to the very general concept of a topological space.

One should be careful that, in general, the family of open sets is not closed under infinite intersections. For example, in $\mathbb{R}$ under the metric $|x - y|$, letting $U_n = (-1/n, +1/n)$, each $U_n$ is open, but $\bigcap_n U_n = \{0\}$, which is not open.

Later on, given any nonempty subset $A$ of a metric space $(E,d)$, we will need to know that certain special sets containing $A$ are open.

**Definition 13.5.** Let $(E,d)$ be a metric space. For any nonempty subset $A$ of $E$ and any $x \in E$, let

$$d(x, A) = \inf_{a \in A} d(x, a).$$

**Proposition 13.2.** Let $(E,d)$ be a metric space. For any nonempty subset $A$ of $E$ and for any two points $x, y \in E$, we have

$$|d(x, A) - d(y, A)| \leq d(x, y).$$
Proof. For all $a \in A$ we have
\[ d(x, a) \leq d(x, y) + d(y, a), \]
which implies
\[
\begin{align*}
    d(x, A) &= \inf_{a \in A} d(x, a) \\
             &\leq \inf_{a \in A} (d(x, y) + d(y, a)) \\
             &= d(x, y) + \inf_{a \in A} d(y, a) \\
             &= d(x, y) + d(y, A).
\end{align*}
\]
By symmetry, we also obtain $d(y, A) \leq d(x, y) + d(x, A)$, and thus
\[
|d(x, A) - d(y, A)| \leq d(x, y),
\]
as claimed. \hfill \qed

Definition 13.6. Let $(E, d)$ be a metric space. For any nonempty subset $A$ of $E$, and any $r > 0$, let
\[
V_r(A) = \{ x \in E \mid d(x, A) < r \}.
\]

Proposition 13.3. Let $(E, d)$ be a metric space. For any nonempty subset $A$ of $E$, and any $r > 0$, the set $V_r(A)$ is an open set containing $A$.

Proof. For any $y \in E$ such that $d(x, y) < r - d(x, A)$, by Proposition 13.2 we have
\[
d(y, A) \leq d(x, A) + d(x, y) \leq d(x, A) + r - d(x, A) = r,
\]
so $V_r(A)$ contains the open ball $B_0(x, r - d(x, A))$, which means that it is open. Obviously, $A \subseteq V_r(A)$. \hfill \qed

13.2 Topological Spaces

Motivated by Proposition 13.1, a topological space is defined in terms of a family of sets satisfying the properties of open sets stated in that proposition.

Definition 13.7. Given a set $E$, a topology on $E$ (or a topological structure on $E$), is defined as a family $\mathcal{O}$ of subsets of $E$ called open sets, and satisfying the following three properties:

(1) For every finite family $(U_i)_{1 \leq i \leq n}$ of sets $U_i \in \mathcal{O}$, we have $U_1 \cap \cdots \cap U_n \in \mathcal{O}$, i.e., $\mathcal{O}$ is closed under finite intersections.

(2) For every arbitrary family $(U_i)_{i \in I}$ of sets $U_i \in \mathcal{O}$, we have $\bigcup_{i \in I} U_i \in \mathcal{O}$, i.e., $\mathcal{O}$ is closed under arbitrary unions.
13.2. TOPOLOGICAL SPACES

(3) $\emptyset \in \mathcal{O}$, and $E \in \mathcal{O}$, i.e., $\emptyset$ and $E$ belong to $\mathcal{O}$.

A set $E$ together with a topology $\mathcal{O}$ on $E$ is called a topological space. Given a topological space $(E, \mathcal{O})$, a subset $F$ of $E$ is a closed set if $F = E - U$ for some open set $U \in \mathcal{O}$, i.e., $F$ is the complement of some open set.

It is possible that an open set is also a closed set. For example, $\emptyset$ and $E$ are both open and closed. When a topological space contains a proper nonempty subset $U$ which is both open and closed, the space $E$ is said to be disconnected.

**Definition 13.8.** A topological space $(E, \mathcal{O})$ is said to satisfy the Hausdorff separation axiom (or $T_2$-separation axiom) if for any two distinct points $a \neq b$ in $E$, there exist two open sets $U_a$ and $U_b$ such that, $a \in U_a$, $b \in U_b$, and $U_a \cap U_b = \emptyset$. When the $T_2$-separation axiom is satisfied, we also say that $(E, \mathcal{O})$ is a Hausdorff space.

As shown by Proposition 13.1, any metric space is a topological Hausdorff space, the family of open sets being in fact the family of arbitrary unions of open balls. Similarly, any normed vector space is a topological Hausdorff space, the family of open sets being the family of arbitrary unions of open balls. The topology $\mathcal{O}$ consisting of all subsets of $E$ is called the discrete topology.

**Remark:** Most (if not all) spaces used in analysis are Hausdorff spaces. Intuitively, the Hausdorff separation axiom says that there are enough “small” open sets. Without this axiom, some counter-intuitive behaviors may arise. For example, a sequence may have more than one limit point (or a compact set may not be closed). Nevertheless, non-Hausdorff topological spaces arise naturally in algebraic geometry. But even there, some substitute for separation is used.

One of the reasons why topological spaces are important is that the definition of a topology only involves a certain family $\mathcal{O}$ of sets, and not how such family is generated from a metric or a norm. For example, different metrics or different norms can define the same family of open sets. Many topological properties only depend on the family $\mathcal{O}$ and not on the specific metric or norm. But the fact that a topology is definable from a metric or a norm is important, because it usually implies nice properties of a space. All our examples will be spaces whose topology is defined by a metric or a norm.

**Definition 13.9.** A topological space $(E, \mathcal{O})$ is metrizable if there is a distance on $E$ defining the topology $\mathcal{O}$.

Note that in a metric space $(E, d)$, the metric $d$ is explicitly given. However, in general, the topology of a metrizable space $(E, \mathcal{O})$ is not specified by an explicitly given metric, but some metric defining the topology $\mathcal{O}$ exists. Obviously, a metrizable topological space must be Hausdorff. Actually, a stronger separation property holds, a metrizable space is normal; see Definition 13.30.
Remark: By taking complements we can state properties of the closed sets dual to those of Definition 13.7. Thus, \( \emptyset \) and \( E \) are closed sets, and the closed sets are closed under finite unions and arbitrary intersections.

It is also worth noting that the Hausdorff separation axiom implies that for every \( a \in E \), the set \( \{a\} \) is closed. Indeed, if \( x \in E - \{a\} \), then \( x \neq a \), and so there exist open sets \( U_a \) and \( U_x \) such that \( a \in U_a \), \( x \in U_x \), and \( U_a \cap U_x = \emptyset \). See Figure 13.6. Thus, for every \( x \in E - \{a\} \), there is an open set \( U_x \) containing \( x \) and contained in \( E - \{a\} \), showing by (O3) that \( E - \{a\} \) is open, and thus that the set \( \{a\} \) is closed.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{hausdorff.png}
\caption{A schematic illustration of the Hausdorff separation property}
\end{figure}

Given a topological space \((E, \mathcal{O})\), given any subset \( A \) of \( E \), since \( E \in \mathcal{O} \) and \( E \) is a closed set, the family \( \mathcal{C}_A = \{F \mid A \subseteq F, F \text{ a closed set}\} \) of closed sets containing \( A \) is nonempty, and since any arbitrary intersection of closed sets is a closed set, the intersection \( \bigcap \mathcal{C}_A \) of the sets in the family \( \mathcal{C}_A \) is the smallest closed set containing \( A \). By a similar reasoning, the union of all the open subsets contained in \( A \) is the largest open set contained in \( A \).

**Definition 13.10.** Given a topological space \((E, \mathcal{O})\), given any subset \( A \) of \( E \), the smallest closed set containing \( A \) is denoted by \( \overline{A} \), and is called the closure, or adherence of \( A \). See Figure 13.7. A subset \( A \) of \( E \) is **dense in** \( E \) if \( \overline{A} = E \). The largest open set contained in \( A \) is denoted by \( \overset{\circ}{A} \), and is called the interior of \( A \). See Figure 13.8. The set \( \text{Fr} \ A = \overline{A} \cap E - A \) is called the boundary (or frontier) of \( A \). We also denote the boundary of \( A \) by \( \partial A \). See Figure 13.9.
Figure 13.7: The topological space \((E, \mathcal{O})\) is \(\mathbb{R}^2\) with topology induced by the Euclidean metric. The subset \(A\) is the section \(B_0(1)\) in the first and fourth quadrants bound by the lines \(y = x\) and \(y = -x\). The closure of \(A\) is obtained by the intersection of \(A\) with the closed unit ball.

Figure 13.8: The topological space \((E, \mathcal{O})\) is \(\mathbb{R}^2\) with topology induced by the Euclidean metric. The subset \(A\) is the section \(B_0(1)\) in the first and fourth quadrants bound by the lines \(y = x\) and \(y = -x\). The interior of \(A\) is obtained by the covering \(A\) with small open balls.
CHAPTER 13. TOPOLOGY

Figure 13.9: The topological space \((E, \mathcal{O})\) is \(\mathbb{R}^2\) with topology induced by the Euclidean metric. The subset \(A\) is the section \(B_0(1)\) in the first and fourth quadrants bound by the lines \(y = x\) and \(y = -x\). The boundary of \(A\) is \(\overline{A} - \mathring{A}\).

Remark: The notation \(\overline{A}\) for the closure of a subset \(A\) of \(E\) is somewhat unfortunate, since \(\overline{A}\) is often used to denote the set complement of \(A\) in \(E\). Still, we prefer it to more cumbersome notations such as \(\text{clo}(A)\), and we denote the complement of \(A\) in \(E\) by \(E - A\) (or sometimes, \(A^c\)).

By definition, it is clear that a subset \(A\) of \(E\) is closed iff \(A = \overline{A}\). The set \(\mathbb{Q}\) of rationals is dense in \(\mathbb{R}\). It is easily shown that \(\overline{A} = \mathring{A} \cup \partial A\) and \(\mathring{A} \cap \partial A = \emptyset\). Another useful characterization of \(\overline{A}\) is given by the following proposition.

**Proposition 13.4.** Given a topological space \((E, \mathcal{O})\), given any subset \(A\) of \(E\), the closure \(\overline{A}\) of \(A\) is the set of all points \(x \in E\) such that for every open set \(U\) containing \(x\), then \(U \cap A \neq \emptyset\). See Figure 13.10.

**Proof.** If \(A = \emptyset\), since \(\emptyset\) is closed, the proposition holds trivially. Thus, assume that \(A \neq \emptyset\). First assume that \(x \in \overline{A}\). Let \(U\) be any open set such that \(x \in U\). If \(U \cap A = \emptyset\), since \(U\) is open, then \(E - U\) is a closed set containing \(A\), and since \(\overline{A}\) is the intersection of all closed sets containing \(A\), we must have \(x \in E - U\), which is impossible. Conversely, assume that \(x \in E\) is a point such that for every open set \(U\) containing \(x\), then \(U \cap A \neq \emptyset\). Let \(F\) be any closed subset containing \(A\). If \(x \notin F\), since \(F\) is closed, then \(U = E - F\) is an open set such that \(x \in U\), and \(U \cap A = \emptyset\), a contradiction. Thus, we have \(x \in F\) for every closed set containing \(A\), that is, \(x \in \overline{A}\).

Often it is necessary to consider a subset \(A\) of a topological space \(E\), and to view the subset \(A\) as a topological space. The following proposition shows how to define a topology on a subset.

**Proposition 13.5.** Given a topological space \((E, \mathcal{O})\), given any subset \(A\) of \(E\), let

\[
U = \{U \cap A \mid U \in \mathcal{O}\}
\]

be the family of all subsets of \(A\) obtained as the intersection of any open set in \(\mathcal{O}\) with \(A\). The following properties hold.
Figure 13.10: The topological space \((E, \mathcal{O})\) is \(\mathbb{R}^2\) with topology induced by the Euclidean metric. The purple subset \(A\) is illustrated with three red points, each in its closure since the open ball centered at each point has nontrivial intersection with \(A\).

(1) The space \((A, \mathcal{U})\) is a topological space.

(2) If \(E\) is a metric space with metric \(d\), then the restriction \(d_A: A \times A \to \mathbb{R}_+\) of the metric \(d\) to \(A\) defines a metric space. Furthermore, the topology induced by the metric \(d_A\) agrees with the topology defined by \(\mathcal{U}\), as above.

Proof. Left as an exercise. \qed

Proposition 13.5 suggests the following definition.

**Definition 13.11.** Given a topological space \((E, \mathcal{O})\), given any subset \(A\) of \(E\), the **subspace topology on \(A\)** induced by \(\mathcal{O}\) is the family \(\mathcal{U}\) of open sets defined such that

\[
\mathcal{U} = \{ U \cap A \mid U \in \mathcal{O} \}
\]

is the family of all subsets of \(A\) obtained as the intersection of any open set in \(\mathcal{O}\) with \(A\). We say that \((A, \mathcal{U})\) has the **subspace topology**. If \((E, d)\) is a metric space, the restriction \(d_A: A \times A \to \mathbb{R}_+\) of the metric \(d\) to \(A\) is called the **subspace metric**.

For example, if \(E = \mathbb{R}^n\) and \(d\) is the Euclidean metric, we obtain the subspace topology on the closed \(n\)-cube

\[
\{(x_1, \ldots, x_n) \in E \mid a_i \leq x_i \leq b_i, \ 1 \leq i \leq n \}.
\]

See Figure 13.11,
Figure 13.11: An example of an open set in the subspace topology for \(\{(x, y, z) \in \mathbb{R}^3 \mid -1 \leq x \leq 1, -1 \leq y \leq 1, -1 \leq z \leq 1\}\). The open set is the corner region \(ABCD\) and is obtained by intersecting with the cube \(B_0((1,1,1),1)\).

One should realize that every open set \(U \in \mathcal{O}\) which is entirely contained in \(A\) is also in the family \(\mathcal{U}\), but \(\mathcal{U}\) may contain open sets that are not in \(\mathcal{O}\). For example, if \(E = \mathbb{R}\) with \(|x - y|\), and \(A = [a, b]\), then sets of the form \([a, c)\), with \(a < c < b\) belong to \(\mathcal{U}\), but they are not open sets for \(\mathbb{R}\) under \(|x - y|\). However, there is agreement in the following situation.

**Proposition 13.6.** Given a topological space \((E, \mathcal{O})\), given any subset \(A\) of \(E\), if \(\mathcal{U}\) is the subspace topology, then the following properties hold.

1. If \(A\) is an open set \(A \in \mathcal{O}\), then every open set \(U \in \mathcal{U}\) is an open set \(U \in \mathcal{O}\).

2. If \(A\) is a closed set in \(E\), then every closed set w.r.t. the subspace topology is a closed set w.r.t. \(\mathcal{O}\).

**Proof.** Left as an exercise. \(\square\)

The concept of product topology is also useful. We have the following proposition.
Proposition 13.7. Given $n$ topological spaces $(E_i, \mathcal{O}_i)$, let $\mathcal{B}$ be the family of subsets of $E_1 \times \cdots \times E_n$ defined as follows:

$$\mathcal{B} = \{ U_1 \times \cdots \times U_n \mid U_i \in \mathcal{O}_i, 1 \leq i \leq n \},$$

and let $\mathcal{P}$ be the family consisting of arbitrary unions of sets in $\mathcal{B}$, including $\emptyset$. Then $\mathcal{P}$ is a topology on $E_1 \times \cdots \times E_n$.

Proof. Left as an exercise. \hfill \Box

Definition 13.12. Given $n$ topological spaces $(E_i, \mathcal{O}_i)$, the product topology on $E_1 \times \cdots \times E_n$ is the family $\mathcal{P}$ of subsets of $E_1 \times \cdots \times E_n$ defined as follows: if $\mathcal{B} = \{ U_1 \times \cdots \times U_n \mid U_i \in \mathcal{O}_i, 1 \leq i \leq n \}$, then $\mathcal{P}$ is the family consisting of arbitrary unions of sets in $\mathcal{B}$, including $\emptyset$. See Figure 13.12.

Figure 13.12: Examples of open sets in the product topology for $\mathbb{R}^2$ and $\mathbb{R}^3$ induced by the Euclidean metric.

If each $(E_i, d_{E_i})$ is a metric space, there are three natural metrics that can be defined on $E_1 \times \cdots \times E_n$:

\begin{align*}
    d_1((x_1, \ldots, x_n), (y_1, \ldots, y_n)) &= d_{E_1}(x_1, y_1) + \cdots + d_{E_n}(x_n, y_n), \\
    d_2((x_1, \ldots, x_n), (y_1, \ldots, y_n)) &= ((d_{E_1}(x_1, y_1))^2 + \cdots + (d_{E_n}(x_n, y_n))^2)^{\frac{1}{2}}, \\
    d_\infty((x_1, \ldots, x_n), (y_1, \ldots, y_n)) &= \max\{d_{E_1}(x_1, y_1), \ldots, d_{E_n}(x_n, y_n)\}.
\end{align*}

It is easy to show that

\begin{align*}
    d_\infty((x_1, \ldots, x_n), (y_1, \ldots, y_n)) &\leq d_2((x_1, \ldots, x_n), (y_1, \ldots, y_n)) \leq d_1((x_1, \ldots, x_n), (y_1, \ldots, y_n)) \\
    &\leq nd_\infty((x_1, \ldots, x_n), (y_1, \ldots, y_n)),
\end{align*}
so these distances define the same topology, which is the product topology.

If each $(E_i, \|\|_{E_i})$ is a normed vector space, there are three natural norms that can be defined on $E_1 \times \cdots \times E_n$:

$$\| (x_1, \ldots, x_n) \|_1 = \| x_1 \|_{E_1} + \cdots + \| x_n \|_{E_n},$$

$$\| (x_1, \ldots, x_n) \|_2 = \left( \| x_1 \|_{E_1}^2 + \cdots + \| x_n \|_{E_n}^2 \right)^{\frac{1}{2}},$$

$$\| (x_1, \ldots, x_n) \|_\infty = \max \{ \| x_1 \|_{E_1}, \ldots, \| x_n \|_{E_n} \}.$$

It is easy to show that

$$\| (x_1, \ldots, x_n) \|_\infty \leq \| (x_1, \ldots, x_n) \|_2 \leq \| (x_1, \ldots, x_n) \|_1 \leq n \| (x_1, \ldots, x_n) \|_\infty,$$

so these norms define the same topology, which is the product topology. It can also be verified that when $E_i = \mathbb{R}$, with the standard topology induced by $|x - y|$, the topology product on $\mathbb{R}^n$ is the standard topology induced by the Euclidean norm.

**Definition 13.13.** Two metrics $d$ and $d'$ on a space $E$ are equivalent if they induce the same topology $\mathcal{O}$ on $E$ (i.e., they define the same family $\mathcal{O}$ of open sets). Similarly, two norms $\|\|$ and $\|\|'$ on a space $E$ are equivalent if they induce the same topology $\mathcal{O}$ on $E$.

Given a topological space $(E, \mathcal{O})$, it is often useful, as in Proposition 13.7, to define the topology $\mathcal{O}$ in terms of a subfamily $\mathcal{B}$ of subsets of $E$.

**Definition 13.14.** We say that a family $\mathcal{B}$ of subsets of $E$ is a basis for the topology $\mathcal{O}$, if $\mathcal{B}$ is a subset of $\mathcal{O}$, and if every open set $U$ in $\mathcal{O}$ can be obtained as some union (possibly infinite) of sets in $\mathcal{B}$ (agreeing that the empty union is the empty set).

For example, given any metric space $(E, d)$, $\mathcal{B} = \{ B_0(a, \rho) \mid a \in E, \rho > 0 \}$. In particular, if $d = \|\|_2$, the open intervals form a basis for $\mathbb{R}$, while the open disks form a basis for $\mathbb{R}^2$. The open rectangles also form a basis for $\mathbb{R}^2$ with the standard topology. See Figure 13.13.

It is immediately verified that if a family $\mathcal{B} = (U_i)_{i \in I}$ is a basis for the topology of $(E, \mathcal{O})$, then $E = \bigcup_{i \in I} U_i$, and the intersection of any two sets $U_i, U_j \in \mathcal{B}$ is the union of some sets in the family $\mathcal{B}$ (again, agreeing that the empty union is the empty set). Conversely, a family $\mathcal{B}$ with these properties is the basis of the topology obtained by forming arbitrary unions of sets in $\mathcal{B}$.

**Definition 13.15.** A subbasis for $\mathcal{O}$ is a family $\mathcal{S}$ of subsets of $E$, such that the family $\mathcal{B}$ of all finite intersections of sets in $\mathcal{S}$ (including $E$ itself, in case of the empty intersection) is a basis of $\mathcal{O}$. See Figure 13.13.

The following proposition gives useful criteria for determining whether a family of open subsets is a basis of a topological space.
13.2. TOPOLOGICAL SPACES

Figure 13.13: Figure (i.) shows that the set of infinite open intervals forms a subbasis for $\mathbb{R}$. Figure (ii.) shows that the infinite open strips form a subbasis for $\mathbb{R}^2$.

**Proposition 13.8.** Given a topological space $(E, \mathcal{O})$ and a family $\mathcal{B}$ of open subsets in $\mathcal{O}$ the following properties hold:

1. The family $\mathcal{B}$ is a basis for the topology $\mathcal{O}$ iff for every open set $U \in \mathcal{O}$ and every $x \in U$, there is some $B \in \mathcal{B}$ such that $x \in B$ and $B \subseteq U$. See Figure 13.14.

2. The family $\mathcal{B}$ is a basis for the topology $\mathcal{O}$ iff
   
   (a) For every $x \in E$, there is some $B \in \mathcal{B}$ such that $x \in B$.
   
   (b) For any two open subsets, $B_1, B_2 \in \mathcal{B}$, for every $x \in E$, if $x \in B_1 \cap B_2$, then there is some $B_3 \in \mathcal{B}$ such that $x \in B_3$ and $B_3 \subseteq B_1 \cap B_2$. See Figure 13.15.

Figure 13.14: Given an open subset $U$ of $\mathbb{R}^2$ and $x \in U$, there exists an open ball $B$ containing $x$ with $B \subset U$. There also exists an open rectangle $B_1$ containing $x$ with $B_1 \subset U$.

We now consider the fundamental property of continuity.
13.3 Continuous Functions, Limits

Definition 13.16. Let \((E, \mathcal{O}_E)\) and \((F, \mathcal{O}_F)\) be topological spaces, and let \(f : E \to F\) be a function. For every \(a \in E\), we say that \(f\) is continuous at \(a\), if for every open set \(V \in \mathcal{O}_F\) containing \(f(a)\), there is some open set \(U \in \mathcal{O}_E\) containing \(a\), such that, \(f(U) \subseteq V\). See Figure 13.16. We say that \(f\) is continuous if it is continuous at every \(a \in E\).

Define a neighborhood of \(a \in E\) as any subset \(N\) of \(E\) containing some open set \(O \in \mathcal{O}\) such that \(a \in O\). If \(f\) is continuous at \(a\) and \(N\) is any neighborhood of \(f(a)\), there is some open set \(V \subseteq N\) containing \(f(a)\), and since \(f\) is continuous at \(a\), there is some open set \(U\) containing \(a\), such that \(f(U) \subseteq V\). Since \(V \subseteq N\), the open set \(U\) is a subset of \(f^{-1}(N)\) containing \(a\), and \(f^{-1}(N)\) is a neighborhood of \(a\). Conversely, if \(f^{-1}(N)\) is a neighborhood of \(a\) whenever \(N\) is any neighborhood of \(f(a)\), it is immediate that \(f\) is continuous at \(a\). See Figure 13.17.

It is easy to see that Definition 13.16 is equivalent to the following statements.

Proposition 13.9. Let \((E, \mathcal{O}_E)\) and \((F, \mathcal{O}_F)\) be topological spaces, and let \(f : E \to F\) be a function. For every \(a \in E\), the function \(f\) is continuous at \(a \in E\) iff for every neighborhood \(N\) of \(f(a) \in F\), then \(f^{-1}(N)\) is a neighborhood of \(a\). The function \(f\) is continuous on \(E\) iff \(f^{-1}(V)\) is an open set in \(\mathcal{O}_E\) for every open set \(V \in \mathcal{O}_F\).
If $E$ and $F$ are metric spaces defined by metrics $d_E$ and $d_F$, we can show easily that $f$ is continuous at $a$ iff for every $\epsilon > 0$, there is some $\eta > 0$, such that, for every $x \in E$,

$$\text{if } d_E(a, x) \leq \eta, \text{ then } d_F(f(a), f(x)) \leq \epsilon.$$  

Similarly, if $E$ and $F$ are normed vector spaces defined by norms $\| \|_E$ and $\| \|_F$, we can show easily that $f$ is continuous at $a$ iff for every $\epsilon > 0$, there is some $\eta > 0$, such that, for every $x \in E$,

$$\text{if } \|x - a\|_E \leq \eta, \text{ then } \|f(x) - f(a)\|_F \leq \epsilon.$$  

It is worth noting that continuity is a topological notion, in the sense that equivalent metrics (or equivalent norms) define exactly the same notion of continuity.

**Definition 13.17.** If $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ are topological spaces, and $f : E \to F$ is a function, for every nonempty subset $A \subseteq E$ of $E$, we say that $f$ is continuous on $A$ if the restriction of $f$ to $A$ is continuous with respect to $(A, \mathcal{U})$ and $(F, \mathcal{O}_F)$, where $\mathcal{U}$ is the subspace topology induced by $\mathcal{O}_E$ on $A$.

Given a product $E_1 \times \cdots \times E_n$ of topological spaces, as usual, we let $\pi_i : E_1 \times \cdots \times E_n \to E_i$ be the projection function such that, $\pi_i(x_1, \ldots, x_n) = x_i$. It is immediately verified that each $\pi_i$ is continuous.

Given a topological space $(E, \mathcal{O})$, we say that a point $a \in E$ is isolated if $\{a\}$ is an open set in $\mathcal{O}$. Then if $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ are topological spaces, any function $f : E \to F$ is continuous at every isolated point $a \in E$. In the discrete topology, every point is isolated.

In a nontrivial normed vector space $(E, \| \|)$ (with $E \neq \{0\}$), no point is isolated. To show this, we show that every open ball $B_0(u, \rho_i)$ contains some vectors different from $u$. 

Figure 13.17: A schematic illustration of the neighborhood condition.
Indeed, since $E$ is nontrivial, there is some $v \in E$ such that $v \neq 0$, and thus $\lambda = \|v\| > 0$ (by (N1)). Let

$$w = u + \frac{\rho}{\lambda + 1}v.$$  

Since $v \neq 0$ and $\rho > 0$, we have $w \neq u$. Then,

$$\|w - u\| = \left\| \frac{\rho}{\lambda + 1}v \right\| = \frac{\rho \lambda}{\lambda + 1} < \rho,$$

which shows that $\|w - u\| < \rho$, for $w \neq u$.

The following proposition is easily shown.

**Proposition 13.10.** Given topological spaces $(E, \mathcal{O}_E)$, $(F, \mathcal{O}_F)$, and $(G, \mathcal{O}_G)$, and two functions $f : E \to F$ and $g : F \to G$, if $f$ is continuous at $a \in E$ and $g$ is continuous at $f(a) \in F$, then $g \circ f : E \to G$ is continuous at $a \in E$. Given $n$ topological spaces $(F_i, \mathcal{O}_i)$, for every function $f : E \to F_1 \times \cdots \times F_n$, then $f$ is continuous at $a \in E$ iff every $f_i : E \to F_i$ is continuous at $a$, where $f_i = \pi_i \circ f$.

One can also show that in a metric space $(E, d)$, the distance $d : E \times E \to \mathbb{R}$ is continuous, where $E \times E$ has the product topology. By the triangle inequality, we have

$$d(x, y) \leq d(x, x_0) + d(x_0, y_0) + d(y_0, y) = d(x_0, y_0) + d(x_0, x) + d(y_0, y)$$

and

$$d(x_0, y_0) \leq d(x_0, x) + d(x, y) + d(y, y_0) = d(x, y) + d(x_0, x) + d(y_0, y).$$

Consequently,

$$|d(x, y) - d(x_0, y_0)| \leq d(x_0, x) + d(y_0, y),$$

which proves that $d$ is continuous at $(x_0, y_0)$. In fact this shows that $d$ is uniformly continuous; see Definition 13.37.

Given any nonempty subset $A$ of $E$, by Proposition 13.2, the map $x \mapsto d(x, A)$ is continuous (in fact, uniformly continuous).

Similarly, for a normed vector space $(E, \|\|)$, the norm $\|\| : E \to \mathbb{R}$ is (uniformly) continuous.

Given a function $f : E_1 \times \cdots \times E_n \to F$, we can fix $n - 1$ of the arguments, say $a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n$, and view $f$ as a function of the remaining argument,

$$x_i \mapsto f(a_1, \ldots, a_{i-1}, x_i, a_{i+1}, \ldots, a_n),$$

where $x_i \in E_i$. If $f$ is continuous, it is clear that each $f_i$ is continuous.
One should be careful that the converse is false! For example, consider the function 
\( f: \mathbb{R} \times \mathbb{R} \to \mathbb{R} \), defined such that,
\[ f(x, y) = \frac{xy}{x^2 + y^2} \text{ if } (x, y) \neq (0, 0), \text{ and } f(0, 0) = 0. \]

The function \( f \) is continuous on \( \mathbb{R} \times \mathbb{R} - \{(0, 0)\} \), but on the line \( y = mx \), with \( m \neq 0 \), we have \( f(x, y) = \frac{m}{1 + m^2} \neq 0 \), and thus, on this line, \( f(x, y) \) does not approach 0 when \((x, y)\) approaches \((0, 0)\). See Figure 13.18.

Figure 13.18: The graph of \( f(x, y) = \frac{xy}{x^2 + y^2} \) for \((x, y) \neq (0, 0)\). The bottom of this graph, which shows the approach along the line \( y = -x \), does not have a \( z \) value of 0.

The following proposition is useful for showing that real-valued functions are continuous.

**Proposition 13.11.** If \( E \) is a topological space, and \((\mathbb{R}, |x - y|)\) the reals under the standard topology, for any two functions \( f: E \to \mathbb{R} \) and \( g: E \to \mathbb{R} \), for any \( a \in E \), for any \( \lambda \in \mathbb{R} \), if \( f \) and \( g \) are continuous at \( a \), then \( f + g, \lambda f, f \cdot g \), are continuous at \( a \), and \( f/g \) is continuous at \( a \) if \( g(a) \neq 0 \).

**Proof.** Left as an exercise. \( \square \)

Using Proposition 13.11, we can show easily that every real polynomial function is continuous.

The notion of isomorphism of topological spaces is defined as follows.

**Definition 13.18.** Let \((E, \mathcal{O}_E)\) and \((F, \mathcal{O}_F)\) be topological spaces, and let \( f: E \to F \) be a function. We say that \( f \) is a homeomorphism between \( E \) and \( F \) if \( f \) is bijective, and both \( f: E \to F \) and \( f^{-1}: F \to E \) are continuous.
One should be careful that a bijective continuous function \( f : E \to F \) is not necessarily a homeomorphism. For example, if \( E = \mathbb{R} \) with the discrete topology, and \( F = \mathbb{R} \) with the standard topology, the identity is not a homeomorphism. Another interesting example involving a parametric curve is given below. Let \( L : \mathbb{R} \to \mathbb{R}^2 \) be the function, defined such that,

\[
L_1(t) = \frac{t(1 + t^2)}{1 + t^4}, \\
L_2(t) = \frac{t(1 - t^2)}{1 + t^4}.
\]

If we think of \((x(t), y(t)) = (L_1(t), L_2(t))\) as a geometric point in \( \mathbb{R}^2 \), the set of points \((x(t), y(t))\) obtained by letting \( t \) vary in \( \mathbb{R} \) from \(-\infty\) to \(+\infty\), defines a curve having the shape of a “figure eight,” with self-intersection at the origin, called the “lemniscate of Bernoulli.” See Figure 13.19. The map \( L \) is continuous, and in fact bijective, but its inverse \( L^{-1} \) is not continuous. Indeed, when we approach the origin on the branch of the curve in the upper left quadrant (i.e., points such that, \( x \leq 0, y \geq 0 \)), then \( t \) goes to \(-\infty\), and when we approach the origin on the branch of the curve in the lower right quadrant (i.e., points such that, \( x \geq 0, y \leq 0 \)), then \( t \) goes to \(+\infty\).

![Figure 13.19: The lemniscate of Bernoulli.](image)

We also review the concept of limit of a sequence. Given any set \( E \), a sequence is any function \( x : \mathbb{N} \to E \), usually denoted by \((x_n)_{n \in \mathbb{N}}\), or \((x_n)_{n \geq 0}\), or even by \((x_n)\).

**Definition 13.19.** Given a topological space \((E, \mathcal{O})\), we say that a sequence \((x_n)_{n \in \mathbb{N}}\) converges to some \( a \in E \) if for every open set \( U \) containing \( a \), there is some \( n_0 \geq 0 \), such that, \( x_n \in U \), for all \( n \geq n_0 \). We also say that \( a \) is a limit of \((x_n)_{n \in \mathbb{N}}\). See Figure 13.20.

When \( E \) is a metric space with metric \( d \), it is easy to show that this is equivalent to the fact that,

for every \( \epsilon > 0 \), there is some \( n_0 \geq 0 \), such that, \( d(x_n, a) \leq \epsilon \), for all \( n \geq n_0 \).

When \( E \) is a normed vector space with norm \( \| \| \), it is easy to show that this is equivalent to the fact that,

for every \( \epsilon > 0 \), there is some \( n_0 \geq 0 \), such that, \( \| x_n - a \| \leq \epsilon \), for all \( n \geq n_0 \).

The following proposition shows the importance of the Hausdorff separation axiom.
Figure 13.20: A schematic illustration of Definition 13.19.

Proposition 13.12. Given a topological space \((E, \mathcal{O})\), if the Hausdorff separation axiom holds, then every sequence has at most one limit.

Proof. Left as an exercise. \(\square\)

It is worth noting that the notion of limit is topological, in the sense that a sequence converge to a limit \(b\) iff it converges to the same limit \(b\) in any equivalent metric (and similarly for equivalent norms).

If \(E\) is a metric space and if \(A\) is a subset of \(E\), there is a convenient way of showing that a point \(x \in E\) belongs to the closure \(\overline{A}\) of \(A\) in terms of sequences.

Proposition 13.13. Given any metric space \((E, d)\), for any subset \(A\) of \(E\) and any point \(x \in E\), we have \(x \in \overline{A}\) iff there is a sequence \((a_n)\) of points \(a_n \in A\) converging to \(x\).

Proof. If the sequence \((a_n)\) of points \(a_n \in A\) converges to \(x\), then for every open subset \(U\) of \(E\) containing \(x\), there is some \(n_0\) such that \(a_n \in U\) for all \(n \geq n_0\), so \(U \cap A \neq \emptyset\), and Proposition 13.4 implies that \(x \in \overline{A}\).

Conversely, assume that \(x \in \overline{A}\). Then for every \(n \geq 1\), consider the open ball \(B_0(x, 1/n)\). By Proposition 13.4, we have \(B_0(x, 1/n) \cap A \neq \emptyset\), so we can pick some \(a_n \in B_0(x, 1/n) \cap A\). This way, we define a sequence \((a_n)\) of points in \(A\), and by construction \(d(x, a_n) < 1/n\) for all \(n \geq 1\), so the sequence \((a_n)\) converges to \(x\). \(\square\)

We still need one more concept of limit for functions.

Definition 13.20. Let \((E, \mathcal{O}_E)\) and \((F, \mathcal{O}_F)\) be topological spaces, let \(A\) be some nonempty subset of \(E\), and let \(f : A \rightarrow F\) be a function. For any \(a \in \overline{A}\) and any \(b \in F\), we say that \(f(x)\) approaches \(b\) as \(x\) approaches \(a\) with values in \(A\) if for every open set \(V \in \mathcal{O}_F\) containing \(b\), there is some open set \(U \in \mathcal{O}_E\) containing \(a\), such that, \(f(U \cap A) \subseteq V\). See Figure 13.21. This is denoted by

\[
\lim_{{x \to a, x \in A}} f(x) = b.
\]
First, note that by Proposition 13.4, since \( a \in \overline{A} \), for every open set \( U \) containing \( a \), we have \( U \cap A \neq \emptyset \), and the definition is nontrivial. Also, even if \( a \in A \), the value \( f(a) \) of \( f \) at \( a \) plays no role in this definition. When \( E \) and \( F \) are metric spaces with metrics \( d_E \) and \( d_F \), it can be shown easily that the definition can be stated as follows:

For every \( \epsilon > 0 \), there is some \( \eta > 0 \), such that, for every \( x \in A \),

\[
\text{if } d_E(x, a) \leq \eta, \text{ then } d_F(f(x), b) \leq \epsilon.
\]

When \( E \) and \( F \) are normed vector spaces with norms \( \| \|_E \) and \( \| \|_F \), it can be shown easily that the definition can be stated as follows:

For every \( \epsilon > 0 \), there is some \( \eta > 0 \), such that, for every \( x \in A \),

\[
\text{if } \|x - a\|_E \leq \eta, \text{ then } \|f(x) - b\|_F \leq \epsilon.
\]

We have the following result relating continuity at a point and the previous notion.

**Proposition 13.14.** Let \((E, \mathcal{O}_E)\) and \((F, \mathcal{O}_F)\) be two topological spaces, and let \( f : E \to F \) be a function. For any \( a \in E \), the function \( f \) is continuous at \( a \) if \( f(x) \) approaches \( f(a) \) when \( x \) approaches \( a \) (with values in \( E \)).

**Proof.** Left as a trivial exercise. \( \square \)

Another important proposition relating the notion of convergence of a sequence to continuity, is stated without proof.

**Proposition 13.15.** Let \((E, \mathcal{O}_E)\) and \((F, \mathcal{O}_F)\) be two topological spaces, and let \( f : E \to F \) be a function.

(1) If \( f \) is continuous, then for every sequence \((x_n)_{n \in \mathbb{N}}\) in \( E \), if \( (x_n) \) converges to \( a \), then \( (f(x_n)) \) converges to \( f(a) \).
13.4. CONNECTED SETS

(2) If $E$ is a metric space, and $(f(x_n))$ converges to $f(a)$ whenever $(x_n)$ converges to $a$, for every sequence $(x_n)_{n \in \mathbb{N}}$ in $E$, then $f$ is continuous.

A special case of Definition 13.20 will be used when $E$ and $F$ are (nontrivial) normed vector spaces with norms $\| \cdot \|_E$ and $\| \cdot \|_F$. Let $U$ be any nonempty open subset of $E$. We showed earlier that $E$ has no isolated points and that every set $\{v\}$ is closed, for every $v \in E$. Since $E$ is nontrivial, for every $v \in U$, there is a nontrivial open ball contained in $U$ (an open ball not reduced to its center). Then, for every $v \in U$, $A = U \setminus \{v\}$ is open and nonempty, and clearly, $v \in \overline{A}$. For any $v \in U$, if $f(x)$ approaches $b$ when $x$ approaches $v$ with values in $A = U \setminus \{v\}$, we say that $f(x)$ approaches $b$ when $x$ approaches $v$ with values $\neq v$ in $U$. This is denoted by

$$\lim_{x \to v, x \in U, x \neq v} f(x) = b.$$ 

Remark: Variations of the above case show up in the following case: $E = \mathbb{R}$, and $F$ is some arbitrary topological space. Let $A$ be some nonempty subset of $\mathbb{R}$, and let $f : A \to F$ be some function. For any $a \in A$, we say that $f$ is continuous on the right at $a$ if

$$\lim_{x \to a, x \in A \cap [a, +\infty)} f(x) = f(a).$$

We can define continuity on the left at $a$ in a similar fashion.

Let us consider another variation. Let $A$ be some nonempty subset of $\mathbb{R}$, and let $f : A \to F$ be some function. For any $a \in A$, we say that $f$ has a discontinuity of the first kind at $a$ if

$$\lim_{x \to a, x \in A \cap (-\infty, a]} f(x) = f(a_-)$$

and

$$\lim_{x \to a, x \in A \cap (a, +\infty)} f(x) = f(a_+),$$

both exist, and either $f(a_-) \neq f(a)$, or $f(a_+) \neq f(a)$.

Note that it is possible that $f(a_-) = f(a_+)$, but $f$ is still discontinuous at $a$ if this common value differs from $f(a)$. Functions defined on a nonempty subset of $\mathbb{R}$, and that are continuous, except for some points of discontinuity of the first kind, play an important role in analysis.

We now turn to connectivity properties of topological spaces.

13.4 Connected Sets

Connectivity properties of topological spaces play a very important role in understanding the topology of surfaces. This section gathers the facts needed to have a good understanding of the classification theorem for compact surfaces (with boundary). The main references are Ahlfors and Sario [1] and Massey [28, 29]. For general background on topology, geometry, and algebraic topology, we also highly recommend Bredon [9] and Fulton [19].
Definition 13.21. A topological space \((E, \mathcal{O})\) is connected if the only subsets of \(E\) that are both open and closed are the empty set and \(E\) itself. Equivalently, \((E, \mathcal{O})\) is connected if \(E\) cannot be written as the union \(E = U \cup V\) of two disjoint nonempty open sets \(U, V\), or if \(E\) cannot be written as the union \(E = U \cup V\) of two disjoint nonempty closed sets. A subset, \(S \subseteq E\), is connected if it is connected in the subspace topology on \(S\) induced by \((E, \mathcal{O})\). See Figure 13.22. A connected open set is called a region and a closed set is a closed region if its interior is a connected (open) set.

Figure 13.22: Figure (i) shows that the union of two disjoint disks in \(\mathbb{R}^2\) is a disconnected set since each circle can be separated by open half regions. Figure (ii) is an example of a connected subset of \(\mathbb{R}^2\) since the two disks can not separated by open sets.

The definition of connectivity is meant to capture the fact that a connected space \(S\) is “one piece.” Given the metric space \((\mathbb{R}^n, \|\|_2)\), the quintessential examples of connected spaces are \(B_0(a, \rho)\) and \(B(a, \rho)\). In particular, the following standard proposition characterizing the connected subsets of \(\mathbb{R}\) can be found in most topology texts (for example, Munkres [33], Schwartz [39]). For the sake of completeness, we give a proof.

Proposition 13.16. A subset of the real line, \(\mathbb{R}\), is connected iff it is an interval, i.e., of the form \([a, b]\), \((a, b]\), where \(a = -\infty\) is possible, \([a, b)\), where \(b = +\infty\) is possible, or \((a, b)\),
where \( a = -\infty \) or \( b = +\infty \) is possible.

**Proof.** Assume that \( A \) is a connected nonempty subset of \( \mathbb{R} \). The cases where \( A = \emptyset \) or \( A \) consists of a single point are trivial. Otherwise, we show that whenever \( a, b \in A \), \( a < b \), then the entire interval \([a, b]\) is a subset of \( A \). Indeed, if this was not the case, there would be some \( c \in (a, b) \) such that \( c \notin A \), and then we could write \( A = ((-\infty, c] \cap A) \cup ([c, +\infty) \cap A) \), where \((-\infty, c] \cap A \) and \((c, +\infty) \cap A \) are nonempty and disjoint open subsets of \( A \), contradicting the fact that \( A \) is connected. It follows easily that \( A \) must be an interval.

Conversely, we show that an interval, \( I \), must be connected. Let \( A \) be any nonempty subset of \( I \) which is both open and closed in \( I \). We show that \( I = A \). Fix any \( x \in A \) and consider the set, \( R_x \), of all \( y \) such that \([x, y] \subseteq A \). If the set \( R_x \) is unbounded, then \( R_x = [x, +\infty) \). Otherwise, if this set is bounded, let \( b \) be its least upper bound. We claim that \( b \) is the right boundary of the interval \( I \). Because \( A \) is closed in \( I \), unless \( I \) is open on the right and \( b \) is its right boundary, we must have \( b \in A \). In the first case, \( A \cap [x, b] = I \cap [x, b] = [x, b] \). In the second case, because \( A \) is also open in \( I \), unless \( b \) is the right boundary of the interval \( I \) (closed on the right), there is some open set \((b - \eta, b + \eta) \) contained in \( A \), which implies that \([x, b + \eta / 2] \subseteq A \), contradicting the fact that \( b \) is the least upper bound of the set \( R_x \). Thus, \( b \) must be the right boundary of the interval \( I \) (closed on the right). A similar argument applies to the set, \( L_y \), of all \( x \) such that \([x, y] \subseteq A \) and either \( L_y \) is unbounded, or its greatest lower bound \( a \) is the left boundary of \( I \) (open or closed on the left). In all cases, we showed that \( A = I \), and the interval must be connected.

Intuitively, if a space is not connected, it is possible to define a continuous function which is constant on disjoint “connected components” and which takes possibly distinct values on disjoint components. This can be stated in terms of the concept of a locally constant function.

**Definition 13.22.** Given two topological spaces \( X, Y \), a function \( f : X \to Y \) is **locally constant** if for every \( x \in X \), there is an open set \( U \subseteq X \) such that \( x \in U \) and \( f \) is constant on \( U \).

We claim that a locally constant function is continuous. In fact, we will prove that \( f^{-1}(V) \) is open for every subset, \( V \subseteq Y \) (not just for an open set \( V \)). It is enough to show that \( f^{-1}(y) \) is open for every \( y \in Y \), since for every subset \( V \subseteq Y \),

\[
    f^{-1}(V) = \bigcup_{y \in V} f^{-1}(y),
\]

and open sets are closed under arbitrary unions. However, either \( f^{-1}(y) = \emptyset \) if \( y \in Y - f(X) \) or \( f \) is constant on \( U = f^{-1}(y) \) if \( y \in f(X) \) (with value \( y \)), and since \( f \) is locally constant, for every \( x \in U \), there is some open set, \( W \subseteq X \), such that \( x \in W \) and \( f \) is constant on \( W \), which implies that \( f(w) = y \) for all \( w \in W \) and thus, that \( W \subseteq U \), showing that \( U \) is a union of open sets and thus, is open. The following proposition shows that a space is connected iff every locally constant function is constant:
CHAPTER 13. TOPOLOGY

Proposition 13.17. A topological space is connected iff every locally constant function is constant. See Figure 13.23.

Proof. First, assume that \( X \) is connected. Let \( f : X \to Y \) be a locally constant function to some space \( Y \) and assume that \( f \) is not constant. Pick any \( y \in f(X) \). Since \( f \) is not constant, \( U_1 = f^{-1}(y) \neq X \), and of course, \( U_1 \neq \emptyset \). We proved just before Proposition 13.17 that \( f^{-1}(V) \) is open for every subset \( V \subseteq Y \), and thus \( U_1 = f^{-1}(y) = f^{-1}(\{y\}) \) and \( U_2 = f^{-1}(Y - \{y\}) \) are both open, nonempty, and clearly \( X = U_1 \cup U_2 \) and \( U_1 \) and \( U_2 \) are disjoint. This contradicts the fact that \( X \) is connected and \( f \) must be constant.

Assume that every locally constant function \( f : X \to Y \) is constant. If \( X \) is not connected, we can write \( X = U_1 \cup U_2 \), where both \( U_1, U_2 \) are open, disjoint, and nonempty. We can define the function, \( f : X \to \mathbb{R} \), such that \( f(x) = 1 \) on \( U_1 \) and \( f(x) = 0 \) on \( U_2 \). Since \( U_1 \) and \( U_2 \) are open, the function \( f \) is locally constant, and yet not constant, a contradiction. \( \square \)

A characterization on the connected subsets of \( \mathbb{R}^n \) is harder and requires the notion of arcwise connectedness. One of the most important properties of connected sets is that they are preserved by continuous maps.

Proposition 13.18. Given any continuous map, \( f : E \to F \), if \( A \subseteq E \) is connected, then \( f(A) \) is connected.

Proof. If \( f(A) \) is not connected, then there exist some nonempty open sets, \( U, V \), in \( F \) such that \( f(A) \cap U \) and \( f(A) \cap V \) are nonempty and disjoint, and

\[
f(A) = (f(A) \cap U) \cup (f(A) \cap V).
\]
Then, $f^{-1}(U)$ and $f^{-1}(V)$ are nonempty and open since $f$ is continuous and
\[ A = (A \cap f^{-1}(U)) \cup (A \cap f^{-1}(V)), \]
with $A \cap f^{-1}(U)$ and $A \cap f^{-1}(V)$ nonempty, disjoint, and open in $A$, contradicting the fact that $A$ is connected.

An important corollary of Proposition 13.18 is that for every continuous function, $f : E \to \mathbb{R}$, where $E$ is a connected space, $f(E)$ is an interval. Indeed, this follows from Proposition 13.16. Thus, if $f$ takes the values $a$ and $b$ where $a < b$, then $f$ takes all values $c \in [a, b]$. This is a very important property known as the intermediate value theorem.

Even if a topological space is not connected, it turns out that it is the disjoint union of maximal connected subsets and these connected components are closed in $E$. In order to obtain this result, we need a few lemmas.

**Lemma 13.19.** Given a topological space, $E$, for any family, $(A_i)_{i \in I}$, of (nonempty) connected subsets of $E$, if $A_i \cap A_j \neq \emptyset$ for all $i, j \in I$, then the union, $A = \bigcup_{i \in I} A_i$, of the family, $(A_i)_{i \in I}$, is also connected.

**Proof.** Assume that $\bigcup_{i \in I} A_i$ is not connected. There exists two nonempty open subsets, $U$ and $V$, of $E$ such that $A \cap U$ and $A \cap V$ are disjoint and nonempty and such that
\[ A = (A \cap U) \cup (A \cap V). \]
Now, for every $i \in I$, we can write
\[ A_i = (A_i \cap U) \cup (A_i \cap V), \]
where $A_i \cap U$ and $A_i \cap V$ are disjoint, since $A_i \subseteq A$ and $A \cap U$ and $A \cap V$ are disjoint. Since $A_i$ is connected, either $A_i \cap U = \emptyset$ or $A_i \cap V = \emptyset$. This implies that either $A_i \subseteq A \cap U$ or $A_i \subseteq A \cap V$. However, by assumption, $A_i \cap A_j \neq \emptyset$, for all $i, j \in I$, and thus, either both $A_i \subseteq A \cap U$ and $A_j \subseteq A \cap U$, or both $A_i \subseteq A \cap V$ and $A_j \subseteq A \cap V$, since $A \cap U$ and $A \cap V$ are disjoint. Thus, we conclude that either $A_i \subseteq A \cap U$ for all $i \in I$, or $A_i \subseteq A \cap V$ for all $i \in I$. But this proves that either
\[ A = \bigcup_{i \in I} A_i \subseteq A \cap U, \]
or
\[ A = \bigcup_{i \in I} A_i \subseteq A \cap V, \]
contradicting the fact that both $A \cap U$ and $A \cap V$ are disjoint and nonempty. Thus, $A$ must be connected. 

In particular, the above lemma applies when the connected sets in a family $(A_i)_{i \in I}$ have a point in common.
Lemma 13.20. If $A$ is a connected subset of a topological space, $E$, then for every subset, $B$, such that $A \subseteq B \subseteq \overline{A}$, where $\overline{A}$ is the closure of $A$ in $E$, the set $B$ is connected.

Proof. If $B$ is not connected, then there are two nonempty open subsets, $U, V$, of $E$ such that $B \cap U$ and $B \cap V$ are disjoint and nonempty, and

$$B = (B \cap U) \cup (B \cap V).$$

Since $A \subseteq B$, the above implies that

$$A = (A \cap U) \cup (A \cap V),$$

and since $A$ is connected, either $A \cap U = \emptyset$, or $A \cap V = \emptyset$. Without loss of generality, assume that $A \cap V = \emptyset$, which implies that $A \subseteq A \cap U \subseteq B \cap U$. However, $B \cap U$ is closed in the subspace topology for $B$ and since $B \subseteq \overline{A}$ and $\overline{A}$ is closed in $E$, the closure of $A$ in $B$ w.r.t. the subspace topology of $B$ is clearly $B \cap \overline{A} = B$, which implies that $B \subseteq B \cap U$ (since the closure is the smallest closed set containing the given set). Thus, $B \cap V = \emptyset$, a contradiction.

In particular, Lemma 13.20 shows that if $A$ is a connected subset, then its closure, $\overline{A}$, is also connected. We are now ready to introduce the connected components of a space.

Definition 13.23. Given a topological space, $(E, O)$, we say that two points, $a, b \in E$, are connected if there is some connected subset, $A$, of $E$ such that $a \in A$ and $b \in A$.

It is immediately verified that the relation “$a$ and $b$ are connected in $E$” is an equivalence relation. Only transitivity is not obvious, but it follows immediately as a special case of Lemma 13.19. Thus, the above equivalence relation defines a partition of $E$ into nonempty disjoint connected components. The following proposition is easily proved using Lemma 13.19 and Lemma 13.20:

Proposition 13.21. Given any topological space, $E$, for any $a \in E$, the connected component containing $a$ is the largest connected set containing $a$. The connected components of $E$ are closed.

The notion of a locally connected space is also useful.

Definition 13.24. A topological space, $(E, O)$, is locally connected if for every $a \in E$, for every neighborhood, $V$, of $a$, there is a connected neighborhood, $U$, of $a$ such that $U \subseteq V$. See Figure 13.24.

As we shall see in a moment, it would be equivalent to require that $E$ has a basis of connected open sets.
13.4. CONNECTED SETS

Figure 13.24: The topological space $E$, which is homeomorphic to an annulus, is locally connected since each point is surrounded by a small disk contained in $E$.

There are connected spaces that are not locally connected and there are locally connected spaces that are not connected. The two properties are independent. For example, the subspace of $\mathbb{R}^2$ $S = \{(x, \sin(1/x)), \mid x > 0\} \cup \{(0, y) \mid -1 \leq y \leq 1\}$ is connected but not locally connected. See Figure 13.25. The subspace $S$ of $\mathbb{R}$ consisting $[0, 1] \cup [2, 3]$ is locally connected but not connected.

**Proposition 13.22.** A topological space, $E$, is locally connected iff for every open subset, $A$, of $E$, the connected components of $A$ are open.

*Proof.* Assume that $E$ is locally connected. Let $A$ be any open subset of $E$ and let $C$ be one of the connected components of $A$. For any $a \in C \subseteq A$, there is some connected neighborhood, $U$, of $a$ such that $U \subseteq A$ and since $C$ is a connected component of $A$ containing $a$, we must have $U \subseteq C$. This shows that for every $a \in C$, there is some open subset containing $a$ contained in $C$, so $C$ is open.

Conversely, assume that for every open subset, $A$, of $E$, the connected components of $A$ are open. Then, for every $a \in E$ and every neighborhood, $U$, of $a$, since $U$ contains some open set $A$ containing $a$, the interior, $\overset{\circ}{U}$, of $U$ is an open set containing $a$ and its connected components are open. In particular, the connected component $C$ containing $a$ is a connected open set containing $a$ and contained in $U$. \hfill \square

Proposition 13.22 shows that in a locally connected space, the connected open sets form a basis for the topology. It is easily seen that $\mathbb{R}^n$ is locally connected. Another very important property of surfaces and more generally, manifolds, is to be arcwise connected. The intuition is that any two points can be joined by a continuous arc of curve. This is formalized as follows.

**Definition 13.25.** Given a topological space, $(E, \mathcal{O})$, an arc (or path) is a continuous map, $\gamma: [a, b] \rightarrow E$, where $[a, b]$ is a closed interval of the real line, $\mathbb{R}$. The point $\gamma(a)$ is the initial
Figure 13.25: Let \( S \) be the graph of \( f(x) = \sin(1/x) \) union the \( y \)-axis between \(-1\) and \(1\). This space is connected, but not locally connected.

The point of the arc and the point \( \gamma(b) \) is the terminal point of the arc. We say that \( \gamma \) is an arc joining \( \gamma(a) \) and \( \gamma(b) \). See Figure 13.26. An arc is a closed curve if \( \gamma(a) = \gamma(b) \). The set \( \gamma([a,b]) \) is the trace of the arc \( \gamma \).

Typically, \( a = 0 \) and \( b = 1 \).

One should not confuse an arc, \( \gamma: [a, b] \rightarrow E \), with its trace. For example, \( \gamma \) could be constant, and thus, its trace reduced to a single point.

An arc is a Jordan arc if \( \gamma \) is a homeomorphism onto its trace. An arc, \( \gamma: [a, b] \rightarrow E \), is a Jordan curve if \( \gamma(a) = \gamma(b) \) and \( \gamma \) is injective on \([a, b]\). Since \([a, b]\) is connected, by Proposition 13.18, the trace \( \gamma([a, b]) \) of an arc is a connected subset of \( E \).

Given two arcs \( \gamma: [0, 1] \rightarrow E \) and \( \delta: [0, 1] \rightarrow E \) such that \( \gamma(1) = \delta(0) \), we can form a new arc defined as follows:

**Definition 13.26.** Given two arcs, \( \gamma: [0, 1] \rightarrow E \) and \( \delta: [0, 1] \rightarrow E \), such that \( \gamma(1) = \delta(0) \), we can form their composition (or product), \( \gamma \delta \), defined such that

\[
\gamma \delta(t) = \begin{cases} 
\gamma(2t) & \text{if } 0 \leq t \leq 1/2; \\
\delta(2t-1) & \text{if } 1/2 \leq t \leq 1.
\end{cases}
\]

The inverse, \( \gamma^{-1} \), of the arc, \( \gamma \), is the arc defined such that \( \gamma^{-1}(t) = \gamma(1-t) \), for all \( t \in [0, 1] \).
13.4. CONNECTED SETS

Figure 13.26: Let $E$ be the torus with subspace topology induced from $\mathbb{R}^3$ with red arc $\gamma([a,b])$. The torus is both arcwise connected and locally arcwise connected.

It is trivially verified that Definition 13.26 yields continuous arcs.

**Definition 13.27.** A topological space, $E$, is *arcwise connected* if for any two points, $a, b \in E$, there is an arc, $\gamma: [0,1] \to E$, joining $a$ and $b$, i.e., such that $\gamma(0) = a$ and $\gamma(1) = b$. A topological space, $E$, is *locally arcwise connected* if for every $a \in E$, for every neighborhood, $V$, of $a$, there is an arcwise connected neighborhood, $U$, of $a$ such that $U \subseteq V$. See Figure 13.26.

The space $\mathbb{R}^n$ is locally arcwise connected, since for any open ball, any two points in this ball are joined by a line segment. Manifolds and surfaces are also locally arcwise connected. Proposition 13.18 also applies to arcwise connectedness (this is a simple exercise). The following theorem is crucial to the theory of manifolds and surfaces:

**Theorem 13.23.** If a topological space, $E$, is arcwise connected, then it is connected. If a topological space, $E$, is connected and locally arcwise connected, then $E$ is arcwise connected.

**Proof.** First, assume that $E$ is arcwise connected. Pick any point, $a$, in $E$. Since $E$ is arcwise connected, for every $b \in E$, there is a path, $\gamma_b: [0,1] \to E$, from $a$ to $b$ and so,

$$E = \bigcup_{b \in E} \gamma_b([0,1])$$

a union of connected subsets all containing $a$. By Lemma 13.19, $E$ is connected.

Now assume that $E$ is connected and locally arcwise connected. For any point $a \in E$, let $F_a$ be the set of all points, $b$, such that there is an arc, $\gamma_b: [0,1] \to E$, from $a$ to $b$. Clearly, $F_a$ contains $a$. We show that $F_a$ is both open and closed. For any $b \in F_a$, since $E$ is locally
arcwise connected, there is an arcwise connected neighborhood $U$ containing $b$ (because $E$ is a neighborhood of $b$). Thus, $b$ can be joined to every point $c \in U$ by an arc, and since by the definition of $F_a$, there is an arc from $a$ to $b$, the composition of these two arcs yields an arc from $a$ to $c$, which shows that $c \in F_a$. But then $U \subseteq F_a$ and thus, $F_a$ is open. See Figure 13.27 (i.). Now assume that $b$ is in the complement of $F_a$. As in the previous case, there is some arcwise connected neighborhood $U$ containing $b$. Thus, every point $c \in U$ can be joined to $b$ by an arc. If there was an arc joining $a$ to $c$, we would get an arc from $a$ to $b$, contradicting the fact that $b$ is in the complement of $F_a$. Thus, every point $c \in U$ is in the complement of $F_a$, which shows that $U$ is contained in the complement of $F_a$, and thus, that the complement of $F_a$ is open. See Figure 13.27 (ii.). Consequently, we have shown that $F_a$ is both open and closed and since it is nonempty, we must have $E = F_a$, which shows that $E$ is arcwise connected.

![Figure 13.27: Schematic illustrations of the proof techniques that show $F_a$ is both open and closed.](image)

If $E$ is locally arcwise connected, the above argument shows that the connected components of $E$ are arcwise connected.
It is not true that a connected space is arcwise connected. For example, the space consisting of the graph of the function

\[ f(x) = \sin\left(\frac{1}{x}\right), \]

where \( x > 0 \), together with the portion of the \( y \)-axis, for which \(-1 \leq y \leq 1\), is connected, but not arcwise connected. See Figure 13.25.

A trivial modification of the proof of Theorem 13.23 shows that in a normed vector space, \( E \), a connected open set is arcwise connected by polygonal lines (i.e., arcs consisting of line segments). This is because in every open ball, any two points are connected by a line segment. Furthermore, if \( E \) is finite dimensional, these polygonal lines can be forced to be parallel to basis vectors.

We now consider compactness.

### 13.5 Compact Sets and Locally Compact Spaces

The property of compactness is very important in topology and analysis. We provide a quick review geared towards the study of manifolds, and for details, we refer the reader to Munkres [33], Schwartz [39]. In this section we will need to assume that the topological spaces are Hausdorff spaces. This is not a luxury, as many of the results are false otherwise.

We begin this section by providing the definition of compactness and describing a collection of compact spaces in \( \mathbb{R} \). There are various equivalent ways of defining compactness. For our purposes, the most convenient way involves the notion of open cover.

**Definition 13.28.** Given a topological space \( E \), for any subset \( A \) of \( E \), an open cover \((U_i)_{i \in I}\) of \( A \) is a family of open subsets of \( E \) such that \( A \subseteq \bigcup_{i \in I} U_i \). An open subcover of an open cover \((U_i)_{i \in I}\) of \( A \) is any subfamily \((U_j)_{j \in J}\) which is an open cover of \( A \), with \( J \subseteq I \). An open cover \((U_i)_{i \in I}\) of \( A \) is finite if \( I \) is finite. See Figure 13.28. The topological space \( E \) is compact if it is Hausdorff and for every open cover \((U_i)_{i \in I}\) of \( E \), there is a finite open subcover \((U_j)_{j \in J}\) of \( E \). Given any subset \( A \) of \( E \), we say that \( A \) is compact if it is compact with respect to the subspace topology. We say that \( A \) is relatively compact if its closure \( \overline{A} \) is compact.

It is immediately verified that a subset \( A \) of \( E \) is compact in the subspace topology relative to \( A \) iff for every open cover \((U_i)_{i \in I}\) of \( A \) by open subsets of \( E \), there is a finite open subcover \((U_j)_{j \in J}\) of \( A \). The property that every open cover contains a finite open subcover is often called the Heine-Borel-Lebesgue property. By considering complements, a Hausdorff space is compact iff for every family \((F_i)_{i \in I}\) of closed sets, if \( \bigcap_{i \in I} F_i = \emptyset \), then \( \bigcap_{j \in J} F_j = \emptyset \) for some finite subset \( J \) of \( I \).

Definition 13.28 requires that a compact space be Hausdorff. There are books in which a compact space is not necessarily required to be Hausdorff. Following Schwartz, we prefer calling such a space quasi-compact.
Another equivalent and useful characterization can be given in terms of families having the finite intersection property.

**Definition 13.29.** A family \((F_i)_{i \in I}\) of sets has the finite intersection property if \(\bigcap_{j \in J} F_j \neq \emptyset\) for every finite subset \(J\) of \(I\).

**Proposition 13.24.** A topological Hausdorff space \(E\) is compact iff for every family \((F_i)_{i \in I}\) of closed sets having the finite intersection property, then \(\bigcap_{i \in I} F_i \neq \emptyset\).

**Proof.** If \(E\) is compact and \((F_i)_{i \in I}\) is a family of closed sets having the finite intersection property, then \(\bigcap_{i \in I} F_i\) cannot be empty, since otherwise we would have \(\bigcap_{j \in J} F_j = \emptyset\) for some finite subset \(J\) of \(I\), a contradiction. The converse is equally obvious. \(\square\)

Another useful consequence of compactness is as follows. For any family \((F_i)_{i \in I}\) of closed sets such that \(F_{i+1} \subseteq F_i\) for all \(i \in I\), if \(\bigcap_{i \in I} F_i = \emptyset\), then \(F_i = \emptyset\) for some \(i \in I\). Indeed, there must be some finite subset \(J\) of \(I\) such that \(\bigcap_{j \in J} F_j = \emptyset\), and since \(F_{i+1} \subseteq F_i\) for all \(i \in I\), we must have \(F_j = \emptyset\) for the smallest \(F_j\) in \((F_j)_{j \in J}\). Using this fact, we note that \(\mathbb{R}\) is not compact. Indeed, the family of closed sets, \(([n, +\infty))_{n \geq 0}\), is decreasing and has an empty intersection.

It is immediately verified that every finite union of compact subsets is compact. Similarly, every finite union of relatively compact subsets is relatively compact (use the fact that \(\overline{A \cup B} = \overline{A \cap B}\) ).
Given a metric space, if we define a bounded subset to be a subset that can be enclosed in some closed ball (of finite radius), then any nonbounded subset of a metric space is not compact. However, a closed interval \([a, b]\) of the real line is compact.

**Proposition 13.25.** Every closed interval, \([a, b]\), of the real line is compact.

**Proof.** We proceed by contradiction. Let \((U_i)_{i \in I}\) be any open cover of \([a, b]\) and assume that there is no finite open subcover. Let \(c = (a + b)/2\). If both \([a, c]\) and \([c, b]\) had some finite open subcover, so would \([a, b]\), and thus, either \([a, c]\) does not have any finite subcover, or \([c, b]\) does not have any finite open subcover. Let \([a_1, b_1]\) be such a bad subinterval. The same argument applies and we split \([a_1, b_1]\) into two equal subintervals, one of which must be bad. Thus, having defined \([a_n, b_n]\) of length \((b - a)/2^n\) as an interval having no finite open subcover, splitting \([a_n, b_n]\) into two equal intervals, we know that at least one of the two has no finite open subcover and we denote such a bad interval by \([a_{n+1}, b_{n+1}]\). See Figure 13.29. The sequence \((a_n)\) is nondecreasing and bounded from above by \(b\), and thus, by a fundamental property of the real line, it converges to its least upper bound, \(\alpha\). Similarly, the sequence \((b_n)\) is nonincreasing and bounded from below by \(a\) and thus, it converges to its greatest lower bound, \(\beta\). However, the common limit \(\alpha = \beta\) of the sequences \((a_n)\) and \((b_n)\), must have \(\alpha = \beta\). However, since \([a_n, b_n]\) has length \((b - a)/2^n\), we must have \(\alpha = \beta\). However, the common limit \(\alpha = \beta\) of the sequences \((a_n)\) and \((b_n)\), must belong to some open set, \(U_i\), of \(A\), and since \(U_i\) is open, it must contain some interval \([c, d]\) containing \(\alpha\). Then, because \(\alpha\) is the common limit of the sequences \((a_n)\) and \((b_n)\), there is some \(N\) such that the intervals \([a_n, b_n]\) are all contained in the interval \([c, d]\) for all \(n \geq N\), which contradicts the fact that none of the intervals \([a_n, b_n]\) has a finite open subcover. Thus, \([a, b]\) is indeed compact.

The argument of Proposition 13.25 can be adapted to show that in \(\mathbb{R}^m\), every closed set, \([a_1, b_1] \times \cdots \times [a_m, b_m]\), is compact. At every stage, we need to divide into \(2^m\) subpieces instead of 2.

We next discuss some important properties of compact spaces. We begin with two separations axioms which only hold for Hausdorff spaces:

**Proposition 13.26.** Given a topological Hausdorff space, \(E\), for every compact subset, \(A\), and every point, \(b\), not in \(A\), there exist disjoint open sets, \(U\) and \(V\), such that \(A \subseteq U\) and \(b \in V\). See Figure 13.30. As a consequence, every compact subset is closed.

**Proof.** Since \(E\) is Hausdorff, for every \(a \in A\), there are some disjoint open sets, \(U_a\) and \(V_a\), containing \(a\) and \(b\) respectively. Thus, the family, \((U_a)_{a \in A}\), forms an open cover of \(A\). Since \(A\) is compact there is a finite open subcover, \((U_j)_{j \in J}\), of \(A\), where \(J \subseteq A\), and then \(\bigcup_{j \in J} U_j\) is an open set containing \(A\) disjoint from the open set \(\bigcap_{j \in J} V_j\) containing \(b\). This shows that every point, \(b\), in the complement of \(A\) belongs to some open set in this complement and thus, that the complement is open, i.e., that \(A\) is closed. See Figure 13.31.

Actually, the proof of Proposition 13.26 can be used to show the following useful property:
Figure 13.29: The first four stages of the nested interval construction utilized in the proof of Proposition 13.25.

**Proposition 13.27.** Given a topological Hausdorff space $E$, for every pair of compact disjoint subsets $A$ and $B$, there exist disjoint open sets $U$ and $V$, such that $A \subseteq U$ and $B \subseteq V$.

**Proof.** We repeat the argument of Proposition 13.26 with $B$ playing the role of $b$ and use Proposition 13.26 to find disjoint open sets $U_a$ containing $a \in A$, and $V_a$ containing $B$. □

The following proposition shows that in a compact topological space, every closed set is compact:

**Proposition 13.28.** Given a compact topological space $E$, every closed set is compact.

**Proof.** Since $A$ is closed, $E - A$ is open and from any open cover, $(U_i)_{i \in I}$, of $A$, we can form an open cover of $E$ by adding $E - A$ to $(U_i)_{i \in I}$ and, since $E$ is compact, a finite subcover, $(U_j)_{j \in J} \cup \{E - A\}$, of $E$ can be extracted such that $(U_j)_{j \in J}$ is a finite subcover of $A$. See Figure 13.32. □

**Remark:** Proposition 13.28 also holds for quasi-compact spaces, i.e., the Hausdorff separation property is not needed.

Putting Proposition 13.27 and Proposition 13.28 together, we note that if $X$ is compact, then for every pair of disjoint closed sets $A$ and $B$, there exist disjoint open sets $U$ and $V$ such that $A \subseteq U$ and $B \subseteq V$. 

13.5. COMPACT SETS AND LOCALLY COMPACT SPACES

Figure 13.30: The compact set of $\mathbb{R}^2$, $A$, is separated by any point in its complement.

Figure 13.31: For the pink compact set $A$, $U$ is the union of the seven disks which cover $A$, while $V$ is the intersection of the seven open sets containing $b$.

**Definition 13.30.** A topological space $E$ is normal if every one-point set is closed, and for every pair of disjoint closed sets $A$ and $B$, there exist disjoint open sets $U$ and $V$ such that $A \subseteq U$ and $B \subseteq V$. A topological space $E$ is regular if every one-point set is closed, and for every point $a \in E$ and every closed subset $B$ of $E$, if $a \notin B$, then there exist disjoint open sets $U$ and $V$ such that $a \in U$ and $B \subseteq V$.

It is clear that a normal space is regular, and a regular space is Hausdorff. There are examples of Hausdorff spaces that are not regular, and of regular spaces that are not normal.

We just observed that a compact space is normal, and this is worth recording as a proposition.

**Proposition 13.29.** Every (Hausdorff) compact space is normal.

An important property of metrizable spaces is that they are normal.

**Proposition 13.30.** Every metrizable space $E$ is normal.
CHAPTER 13. TOPOLOGY

Figure 13.32: An illustration of the proof of Proposition 13.28. Both $E$ and $A$ are closed squares in $\mathbb{R}^2$. Note that an open cover of $A$, namely the green circles, when combined with the yellow square annulus $E - A$ covers all of the yellow square $E$.

Proof. Assume the topology of $E$ is given by the metric $d$. Since $B$ is closed and $A \cap B = \emptyset$, for every $a \in A$ since $a \notin B = B$, there is some open ball $B_0(a, \epsilon_a)$ of radius $\epsilon_a > 0$ such that $B_0(a, \epsilon_a) \cap B = \emptyset$. Similarly, since $A$ is closed and $A \cap B = \emptyset$, for every $b \in B$ there is some open ball $B_0(b, \epsilon_b)$ of radius $\epsilon_b > 0$ such that $B_0(b, \epsilon_b) \cap A = \emptyset$. Let

$$U = \bigcup_{a \in A} B_0(a, \epsilon_a/2), \quad V = \bigcup_{b \in B} B_0(b, \epsilon_b/2).$$

Then $A$ and $B$ are open sets such that $A \subseteq U$ and $B \subseteq V$, and we claim that $U \cap V = \emptyset$.

If not, then there is some $z \in U \cap V$, which implies that for some $a \in A$ and some $b \in B$, we have

$$z \in B_0(a, \epsilon_a/2) \cap B_0(b, \epsilon_b/2).$$

It follows that

$$d(a, b) \leq d(a, z) + d(z, b) < (\epsilon_a + \epsilon_b)/2.$$ 

If $\epsilon_a \leq \epsilon_b$, then $d(a, b) < \epsilon_b$, so $a \in B_0(b, \epsilon_b)$, contradicting the fact that $B_0(b, \epsilon_b) \cap A = \emptyset$. If $\epsilon_b \leq \epsilon_a$, then $d(a, b) < \epsilon_a$, so $b \in B_0(a, \epsilon_a)$, contradicting the fact that $B_0(a, \epsilon_a) \cap B = \emptyset$. $\square$
Normal spaces have a strong separation property regarding disjoint closed subsets $A$ and $B$. Actually, this separation property can be stated as the existence of a certain continuous function $f: E \rightarrow [0,1]$ taking the value 1 on $A$ and the value 0 on $B$. This result is known as Urysohn Lemma. It is an important tool in topology and analysis.

**Theorem 13.31. (Urysohn Lemma)** Let $E$ be a normal space. For any two closed disjoint subsets $A$ and $B$, there is a continuous function $f: E \rightarrow [0,1]$ such that $f(x) = 1$ for all $x \in A$ and $f(x) = 0$ for all $x \in B$.

A proof of Theorem 13.31 can be found in Munkres [33] (Chapter 4, Section 33, Theorem 33.1). Theorem 13.31 is one of the ingredients in the Urysohn metrization theorem (Theorem 13.44).

Compact spaces also have the following property.

**Proposition 13.32.** Given a compact topological space, $E$, for every $a \in E$, for every neighborhood, $V$, of $a$, there exists a compact neighborhood, $U$, of $a$ such that $U \subseteq V$. See Figure 13.33.

![Figure 13.33](image)

Figure 13.33: Let $E$ be the peach square of $\mathbb{R}^2$. Each point of $E$ is contained in a compact neighborhood $U$, in this case the small closed yellow disk.

**Proof.** Since $V$ is a neighborhood of $a$, there is some open subset, $O$, of $V$ containing $a$. Then the complement, $K = E - O$, of $O$ is closed and since $E$ is compact, by Proposition 13.28, $K$ is compact. Now, if we consider the family of all closed sets of the form, $K \cap F$, where $F$ is any closed neighborhood of $a$, since $a \notin K$, this family has an empty intersection and thus, there is a finite number of closed neighborhood, $F_1, \ldots, F_n$, of $a$, such that $K \cap F_1 \cap \cdots \cap F_n = \emptyset$. Then, $U = F_1 \cap \cdots \cap F_n$ is closed and hence by Proposition 13.28, a compact neighborhood of $a$ contained in $O \subseteq V$. See Figure 13.34. □
Figure 13.34: Let $E$ be the peach square of $\mathbb{R}^2$. The compact neighborhood of $a$, $U$, is the intersection of the closed sets $F_1, F_2, F_3$, each of which are contained in the complement of $K$.

It can be shown that in a normed vector space of finite dimension, a subset is compact iff it is closed and bounded. For $\mathbb{R}^n$ the proof is simple.

In a normed vector space of infinite dimension, there are closed and bounded sets that are not compact!

More could be said about compactness in metric spaces but we will only need the notion of Lebesgue number, which will be discussed a little later. Another crucial property of compactness is that it is preserved under continuity.

**Proposition 13.33.** Let $E$ be a topological space and let $F$ be a topological Hausdorff space. For every compact subset, $A$, of $E$, for every continuous map, $f: E \rightarrow F$, the subspace $f(A)$ is compact.

**Proof.** Let $(U_i)_{i \in I}$ be an open cover of $f(A)$. We claim that $(f^{-1}(U_i))_{i \in I}$ is an open cover of $A$, which is easily checked. Since $A$ is compact, there is a finite open subcover, $(f^{-1}(U_j))_{j \in J}$, of $A$, and thus, $(U_j)_{j \in J}$ is an open subcover of $f(A)$. \qed
As a corollary of Proposition 13.33, if $E$ is compact, $F$ is Hausdorff, and $f : E \to F$ is continuous and bijective, then $f$ is a homeomorphism. Indeed, it is enough to show that $f^{-1}$ is continuous, which is equivalent to showing that $f$ maps closed sets to closed sets. However, closed sets are compact and Proposition 13.33 shows that compact sets are mapped to compact sets, which, by Proposition 13.26, are closed.

Another important corollary of Proposition 13.33 is the following result.

**Proposition 13.34.** If $E$ is a compact nonempty topological space and if $f : E \to \mathbb{R}$ is a continuous function, then there are points $a, b \in E$ such that $f(a)$ is the minimum of $f(E)$ and $f(b)$ is the maximum of $f(E)$.

**Proof.** The set $f(E)$ is a compact subset of $\mathbb{R}$ and thus, a closed and bounded set which contains its greatest lower bound and its least upper bound. \hfill \Box

The following property also holds.

**Proposition 13.35.** Let $(E, d)$ be a metric space. For any nonempty subset $A$ of $E$, if $A$ is compact, then for every open subset $U$ such that $A \subseteq U$, there is some $r > 0$ such that $V_r(A) \subseteq U$.

**Proof.** The function $x \mapsto d(x, E - U)$ is continuous and $d(x, E - U) > 0$ for $x \in A$ (since $A \subseteq U$). By Proposition 13.34, there is some $a \in A$ such that

$$d(a, E - U) = \inf_{x \in A} d(x, E - U).$$

But $d(a, E - U) = r > 0$, which implies that $V_r(A) \subseteq U$. \hfill \Box

Another useful notion is that of local compactness. Indeed manifolds and surfaces are locally compact.

**Definition 13.31.** A topological space $E$ is **locally compact** if it is Hausdorff and for every $a \in E$, there is some compact neighborhood $K$ of $a$. See Figure 13.33.

From Proposition 13.32, every compact space is locally compact but the converse is false. For example, $\mathbb{R}$ is locally compact but not compact. In fact it can be shown that a normed vector space of finite dimension is locally compact.

**Proposition 13.36.** Given a locally compact topological space, $E$, for every $a \in E$, for every neighborhood, $N$, of $a$, there exists a compact neighborhood, $U$, of $a$, such that $U \subseteq N$.

**Proof.** For any $a \in E$, there is some compact neighborhood, $V$, of $a$. By Proposition 13.32, every neighborhood of $a$ relative to $V$ contains some compact neighborhood $U$ of $a$ relative to $V$. But every neighborhood of $a$ relative to $V$ is a neighborhood of $a$ relative to $E$ and every neighborhood $N$ of $a$ in $E$ yields a neighborhood, $V \cap N$, of $a$ in $V$ and thus, for every neighborhood, $N$, of $a$, there exists a compact neighborhood, $U$, of $a$ such that $U \subseteq N$. \hfill \Box
When \( E \) is a metric space, the subsets \( V_r(A) \) defined in Definition 13.6 have the following property.

**Proposition 13.37.** Let \((E,d)\) be a metric space. If \( E \) is locally compact, then for any nonempty compact subset \( A \) of \( E \), there is some \( r > 0 \) such that \( V_r(A) \) is compact.

**Proof.** Since \( E \) is locally compact, for every \( x \in A \), there is some compact subset \( V_x \) whose interior \( \text{int} V_x \) contains \( x \). The family of open subsets \( \{V_x\} \) is an open cover of \( A \), and since \( A \) is compact, it has a finite subcover \( \{V_{x_1}, \ldots, V_{x_n}\} \). Then \( U = V_{x_1} \cup \cdots \cup V_{x_n} \) is compact (as a finite union of compact subsets), and it contains an open subset containing \( A \) (the union of the \( V_{x_i} \)). By Proposition 13.35, there is some \( r > 0 \) such that \( V_r(A) \subseteq U \), and thus \( V_r(A) \) is compact. \( \square \)

Another very important property of locally compact spaces is the Proposition 13.39 below. This result implies the existence of continuous partitions of unity for a finite open cover of a compact subset. Such partitions of unity are used in proving that Radon functionals correspond to certain Borel measures. First we have the following proposition.

**Proposition 13.38.** Let \( E \) be a locally compact (Hausdorff) space. For every compact subset \( K \) and every open subset \( V \), if \( K \subseteq V \), then there is an open set \( W \) with compact closure such that \( K \subseteq W \subseteq W \subseteq V \).

A proof of Proposition 13.38 can be found in Rudin [36] (Chapter 2, Theorem 2.7). The following proposition shows the existence of continuous “bump functions” in a locally compact space. It is sometimes called Urysohn lemma (which is a bit confusing since there is already a Urysohn lemma (Proposition 13.31)).

**Proposition 13.39.** Let \( E \) be a locally compact (Hausdorff) space. For every compact subset \( K \) and every open subset \( V \) of \( E \), if \( K \subseteq V \), there is a continuous function \( f: E \to [0,1] \) such that \( f(x) = 1 \) for all \( x \in K \), and such that \( \text{supp}(f) \) is compact and \( \text{supp}(f) \subseteq V \), where \( \text{supp}(f) \) is the closure of the subset \( \{x \in E \mid f(x) \neq 0\} \), called the support of \( f \).

**Proof.** Theorem 13.39 follows easily from the Urysohn lemma (Theorem 13.31). Since \( E \) is locally compact, by Proposition 13.38 we can find some open subset \( W \) with compact closure \( \overline{W} \) such that \( K \subseteq W \subseteq \overline{W} \subseteq V \). Since \( \overline{W} \) is compact, it is normal, so we can apply Theorem 13.31 to find a continuous function \( f: \overline{W} \to [0,1] \) such that \( f(x) = 1 \) for all \( x \in K \) and \( f(x) = 0 \) for all \( x \in \overline{W} - W \) (the boundary of \( W \)). Then we extend \( f \) to \( E \) by setting to 0 outside \( \overline{W} \). Since the support of \( f \) is contained in \( \overline{W} \), this function is continuous. \( \square \)

As a corollary of Proposition 13.39 we obtain the existence of continuous partitions of unity for a finite open cover of a compact subset.
Proposition 13.40. Let $E$ be a locally compact (Hausdorff) space. For any compact subset $K$ of $E$ and any finite open cover $(U_1, \ldots, U_n)$ of $K$ (that is, $K \subseteq \bigcup_{i=1}^{n} U_i$), there exist $n$ continuous functions $f_i: E \to [0,1]$ such that $f_i$ has compact support $\text{supp}(f_i) \subseteq U_i$, and

$$\sum_{i=1}^{n} f_i(x) = 1 \quad \text{for all } x \in K.$$
Figure 13.35: The stereographic projections of $x^2 + y^2 + (z - 1)^2 = 1$ onto the $xy$-plane.

The following theorem shows that $(E_\omega, \mathcal{O}_\omega)$ is indeed a topological space, and that it is compact.

**Theorem 13.41.** Let $E$ be a locally compact topological space. The Alexandroff compactification, $E_\omega$, of $E$ is a compact space such that $E$ is a subspace of $E_\omega$ and if $E$ is not compact, then $\overline{E} = E_\omega$.

**Proof.** The verification that $\mathcal{O}_\omega$ is a family of open sets is not difficult but a bit tedious. Details can be found in Munkres [33] or Schwartz [39]. Let us show that $E_\omega$ is compact. For every open cover, $(U_i)_{i \in I}$, of $E_\omega$, since $\omega$ must be covered, there is some $U_{i_0}$ of the form

$$U_{i_0} = (E - K_0) \cup \{\omega\}$$

where $K_0$ is compact in $E$. Consider the family, $(V_i)_{i \in I}$, defined as follows:

$$V_i = U_i \quad \text{if} \quad U_i \in \mathcal{O},$$

$$V_i = E - K \quad \text{if} \quad U_i = (E - K) \cup \{\omega\},$$

where $K$ is compact in $E$. Then, because each $K$ is compact and thus closed in $E$ (since $E$ is Hausdorff), $E - K$ is open, and every $V_i$ is an open subset of $E$. Furthermore, the family, $(V_i)_{i \in I \setminus \{i_0\}}$, is an open cover of $K_0$. Since $K_0$ is compact, there is a finite open subcover, $(V_j)_{j \in J}$, of $K_0$, and thus, $(U_j)_{j \in J \cup \{i_0\}}$ is a finite open cover of $E_\omega$.

Let us show that $E_\omega$ is Hausdorff. Given any two points, $a, b \in E_\omega$, if both $a, b \in E$, since $E$ is Hausdorff and every open set in $\mathcal{O}$ is an open set in $\mathcal{O}_\omega$, there exist disjoint open sets, $U, V$ (in $\mathcal{O}$), such that $a \in U$ and $b \in V$. If $b = \omega$, since $E$ is locally compact, there is some compact set, $K$, containing an open set, $U$, containing $a$ and then, $U$ and $V = (E - K) \cup \{\omega\}$ are disjoint open sets (in $\mathcal{O}_\omega$) such that $a \in U$ and $b \in V$. 
The space $E$ is a subspace of $E_\omega$ because for every open set, $U$, in $\mathcal{O}_\omega$, either $U \in \mathcal{O}$ and $E \cap U = U$ is open in $E$, or $U = (E - K) \cup \{\omega\}$, where $K$ is compact in $E$, and thus, $U \cap E = E - K$, which is open in $E$, since $K$ is compact in $E$ and thus, closed (since $E$ is Hausdorff). Finally, if $E$ is not compact, for every compact subset, $K$, of $E$, $E - K$ is nonempty and thus, for every open set, $U = (E - K) \cup \{\omega\}$, containing $\omega$, we have $U \cap E \neq \emptyset$, which shows that $\omega \in \overline{E}$ and thus, that $\overline{E} = E_\omega$. \hfill $\square$

13.6 Second-Countable and Separable Spaces

In studying surfaces and manifolds, an important property is the existence of a countable basis for the topology. Indeed this property, among other things, guarantees the existence of triangulations of manifolds, and the fact that a manifold is metrizable.

**Definition 13.33.** A topological space $E$ is called second-countable if there is a countable basis for its topology, i.e., if there is a countable family, $(U_i)_{i \geq 0}$, of open sets such that every open set of $E$ is a union of open sets $U_i$.

It is easily seen that $\mathbb{R}^n$ is second-countable and more generally, that every normed vector space of finite dimension is second-countable. More generally, a metric space is second-countable if and only if it is separable, a very useful property that holds for all of the spaces that we will consider in practice.
Figure 13.37: The two types of open sets associated with the Alexandroff compactification of the $xy$-plane. The first type of open set does not include $\omega$, i.e. the north pole, while the second type of open set contains $\omega$.

**Definition 13.34.** A topological space $E$ is *separable* if it contains some countable subset $S$ which is dense in $X$, that is, $\overline{S} = E$.

Observe that by Proposition 13.4, a subset $S$ of $E$ is dense in $E$ iff every nonempty open subset of $E$ contains some element of $S$.

The (metric) space $\mathbb{R}$ is separable because $\mathbb{Q}$ is a countable dense subset of $\mathbb{R}$. Similarly, $\mathbb{C}$ is separable. In general, $\mathbb{Q}^n$ is dense in $\mathbb{R}^n$, so $\mathbb{R}^n$ is separable, and similarly, every finite-dimensional normed vector space over $\mathbb{R}$ (or $\mathbb{C}$) is separable. For metric spaces, we have the following useful result.

**Proposition 13.42.** If $E$ is a metric space, then $E$ is second-countable if and only if $E$ is separable.

*Proof.* If $\mathcal{B} = (B_n)$ is a countable basis for the topology of $E$, then for any set $S$ obtained by picking some point $s_n$ in $B_n$, since every nonempty open subset $U$ of $E$ is the union of some of the $B_n$, the intersection $U \cap S$ is nonempty, and so $S$ is dense in $E$.

Conversely, assume that there is a countable subset $S = (s_n)$ of $E$ which is dense in $E$. We claim that the countable family $\mathcal{B}$ of open balls $B_{0}(s_n, 1/m)$ ($m \in \mathbb{N}, m > 0$) is a basis for the topology of $E$. For every $x \in E$ and every $r > 0$, there is some $m > 0$ such that
1.3.6. SECOND-COUNTABLE AND SEPARABLE SPACES

1/m < r/2, and some n such that s_n ∈ B_0(x, 1/m). It follows that x ∈ B_0(s_n, 1/m). For all y ∈ B_0(s_n, 1/m), we have
\[d(x, y) ≤ d(x, s_n) + d(s_n, y) ≤ 2/m < r,\]
thus B_0(s_n, 1/m) ⊆ B_0(x, r), which by Proposition 13.8(a) implies that B is a basis for the topology of E.

**Proposition 13.43.** If E is a compact metric space, then E is separable.

*Proof.* For every n > 0, the family of open balls of radius 1/n forms an open cover of E, and since E is compact, there is a finite subset A_n of E such that E = \(\bigcup_{a_i \in A_n} B_0(a_i, 1/n)\). It is easy to see that this is equivalent to the condition d(x, A_n) < 1/n for all x ∈ E. Let A = \(\bigcup_{n \geq 1} A_n\). Then A is countable, and for every x ∈ E, we have
\[d(x, A) ≤ d(x, A_n) < \frac{1}{n}, \text{ for all } n \geq 1,\]
which implies that d(x, A) = 0; that is, A is dense in E. □

The following theorem due to Uryshon gives a very useful sufficient condition for a topological space to be metrizable.

**Theorem 13.44.** (Urysohn metrization theorem) If a topological space E is regular and second-countable, then it is metrizable.

The proof of Theorem 13.44 can be found in Munkres [33] (Chapter 4, Theorem 34.1). As a corollary of Theorem 13.44, every (second-countable) manifold, and thus every Lie group is metrizable.

The following technical result shows that a locally compact metrizable space which is also separable can be expressed as the union of a countable monotonic sequence of compact subsets. This gives us a method for generalizing various properties of compact metric spaces to locally compact metric spaces of the above kind.

**Proposition 13.45.** Let E be a locally compact metrizable space. The following properties are equivalent:

1. There is a sequence \((U_n)_{n \geq 0}\) of open subsets such that for all n ∈ \(\mathbb{N}\), \(U_n \subseteq U_{n+1}\), \(\overline{U_n}\) is compact, \(\overline{U_n} \subseteq U_{n+1}\), and \(E = \bigcup_{n \geq 0} U_n = \bigcup_{n \geq 0} \overline{U_n}\).

2. The space E is the union of a countable family of compact subsets of E.

3. The space E is separable.
Proof. We show (1) implies (2), (2) implies (3), and (3) implies (1). Obviously, (1) implies (2) since the $U_n$ are compact.

If (2) holds, then $E = \bigcup_{n \geq 0} K_n$, for some compact subsets $K_n$. By Proposition 13.43, each compact subset $K_n$ is separable, so let $S_n$ be a countable dense subset of $K_n$. Then $S = \bigcup_{n \geq 0} S_n$ is a countable dense subset of $E$, since

$$E = \bigcup_{n \geq 0} K_n \subseteq \bigcup_{n \geq 0} S_n \subseteq S \subseteq E.$$ 

Consequently (3) holds.

If (3) holds, let $S = \{s_n\}$ be a countable dense subset of $E$. By Proposition 13.42, the space $E$ has a countable basis $B$ of open sets $O_n$. Since $E$ is locally compact, for every $x \in E$, there is some compact neighborhood $W_x$ containing $x$, and by Proposition 13.8, there some index $n(x)$ such that $x \in O_{n(x)} \subseteq W_x$. Since $W_x$ is a compact neighborhood, we deduce that $\overline{O_{n(x)}}$ is compact. Consequently, there is a subfamily of $B$ consisting of open subsets $O_i$ such that $O_i$ is compact, which is a countable basis for the topology of $E$, so we may assume that we restrict our attention to this basis. We define the sequence $(U_n)_{n \geq 1}$ of open subsets of $E$ by induction as follows: Set $U_1 = O_1$, and let

$$U_{n+1} = O_{n+1} \cup V_r(U_n),$$

where $r > 0$ is chosen so that $V_r(U_n)$ is compact, which is possible by Proposition 13.37. We immediately check that the $U_n$ satisfy (1) of Proposition 13.45. □

Definition 13.35. Given a topological space $E$, a subset $A$ of $E$ is $\sigma$-compact (or countable at infinity) if $A$ is the union of countably many compact subsets.

Note that Proposition 13.45 implies that a locally compact metrizable space is separable iff it is $\sigma$-compact.

It can also be shown that if $E$ is a locally compact space that has a countable basis, then $E_\omega$ also has a countable basis (and in fact, is metrizable).

We also have the following property.

Proposition 13.46. Given a second-countable topological space $E$, every open cover $(U_i)_{i \in I}$ of $E$ contains some countable subcover.

Proof. Let $(O_n)_{n \geq 0}$ be a countable basis for the topology. Then all sets $O_n$ contained in some $U_i$ can be arranged into a countable subsequence, $(\Omega_m)_{m \geq 0}$, of $(O_n)_{n \geq 0}$ and for every $\Omega_m$, there is some $U_{i_m}$ such that $\Omega_m \subseteq U_{i_m}$. Furthermore, every $U_i$ is some union of sets $\Omega_j$, and thus, every $a \in E$ belongs to some $\Omega_j$, which shows that $(\Omega_m)_{m \geq 0}$ is a countable open subcover of $(U_i)_{i \in I}$. □

As an immediate corollary of Proposition 13.46, a locally connected second-countable space has countably many connected components.
13.7 Sequential Compactness

For a general topological Hausdorff space $E$, the definition of compactness relies on the existence of finite cover. However, when $E$ has a countable basis or is a metric space, we may define the notion of compactness in terms of sequences. To understand how this is done, we need to first define accumulation points.

**Definition 13.36.** Given a topological Hausdorff space, $E$, given any sequence, $(x_n)$, of points in $E$, a point, $l \in E$, is an accumulation point (or cluster point) of the sequence $(x_n)$ if every open set, $U$, containing $l$ contains $x_n$ for infinitely many $n$. See Figure 13.38.

Figure 13.38: The space $E$ is the closed, bounded pink subset of $\mathbb{R}^2$. The sequence $(x_n)$ has two accumulation points, one for the subsequence $(x_{2n+1})$ and one for $(x_{2n})$.

Clearly, if $l$ is a limit of the sequence, $(x_n)$, then it is an accumulation point, since every open set, $U$, containing $l$ contains all $x_n$ except for finitely many $n$.

For second-countable spaces we are able to give another characterization of accumulation points.

**Proposition 13.47.** Given a second-countable topological Hausdorff space, $E$, a point, $l$, is an accumulation point of the sequence, $(x_n)$, iff $l$ is the limit of some subsequence, $(x_{n_k})$, of $(x_n)$.

**Proof.** Clearly, if $l$ is the limit of some subsequence $(x_{n_k})$ of $(x_n)$, it is an accumulation point of $(x_n)$.

Conversely, let $(U_k)_{k \geq 0}$ be the sequence of open sets containing $l$, where each $U_k$ belongs to a countable basis of $E$, and let $V_k = U_1 \cap \cdots \cap U_k$. For every $k \geq 1$, we can find some $n_k > n_{k-1}$ such that $x_{n_k} \in V_k$, since $l$ is an accumulation point of $(x_n)$. Now, since every open set containing $l$ contains some $U_{k_0}$ and since $x_{n_k} \in U_{k_0}$ for all $k \geq 0$, the sequence $(x_{n_k})$ has limit $l$. 

\hfill $\square$
Remark: Proposition 13.47 also holds for metric spaces.

As an illustration of Proposition 13.47 let \((x_n)\) be the sequence \((1, -1, 1, -1, \ldots)\). This sequence has two accumulation points, namely 1 and \(-1\) since \((x_{2n+1}) = (1)\) and \((x_{2n}) = (-1)\).

In second-countable Hausdorff spaces, compactness can be characterized in terms of accumulation points (this is also true for metric spaces).

**Proposition 13.48.** A second-countable topological Hausdorff space, \(E\), is compact iff every sequence, \((x_n)\), of \(E\) has some accumulation point in \(E\).

**Proof.** Assume that every sequence, \((x_n)\), has some accumulation point. Let \((U_i)_{i \in I}\) be some open cover of \(E\). By Proposition 13.46, there is a countable open subcover, \((O_n)_{n \geq 0}\), for \(E\). Now, if \(E\) is not covered by any finite subcover of \((O_n)_{n \geq 0}\), we can define a sequence, \((x_m)\), by induction as follows:

Let \(x_0\) be arbitrary and for every \(m \geq 1\), let \(x_m\) be some point in \(E\) not in \(O_1 \cup \cdots \cup O_m\), which exists, since \(O_1 \cup \cdots \cup O_m\) is not an open cover of \(E\). We claim that the sequence, \((x_m)\), does not have any accumulation point. Indeed, for every \(l \in E\), since \((O_n)_{n \geq 0}\) is an open cover of \(E\), there is some \(O_m\) such that \(l \in O_m\), and by construction, every \(x_n\) with \(n \geq m + 1\) does not belong to \(O_m\), which means that \(x_n \in O_m\) for only finitely many \(n\) and \(l\) is not an accumulation point. See Figure 13.39.

![Figure 13.39: The space \(E\) is the open half plane above the line \(y = -1\). Since \(E\) is not compact, we inductively build a sequence, \((x_n)\) that will have no accumulation point in \(E\). Note the \(y\) coordinate of \(x_n\) approaches infinity.](image)

Conversely, assume that \(E\) is compact, and let \((x_n)\) be any sequence. If \(l \in E\) is not an accumulation point of the sequence, then there is some open set, \(U_l\), such that \(l \in U_l\).
and $x_n \in U_l$ for only finitely many $n$. Thus, if $(x_n)$ does not have any accumulation point, the family, $(U_l)_{l \in E}$, is an open cover of $E$ and since $E$ is compact, it has some finite open subcover, $(U_l)_{l \in J}$, where $J$ is a finite subset of $E$. But every $U_l$ with $l \in J$ is such that $x_n \in U_l$ for only finitely many $n$, and since $J$ is finite, $x_n \in \bigcup_{l \in J} U_l$ for only finitely many $n$, which contradicts the fact that $(U_l)_{l \in J}$ is an open cover of $E$, and thus contains all the $x_n$. Thus, $(x_n)$ has some accumulation point. See Figure 13.40.

Figure 13.40: The space $E$ the closed triangular region of $\mathbb{R}^2$. Given a sequence $(x_n)$ of red points in $E$, if the sequence has no accumulation points, then each $l_i$ for $1 \leq i \leq 8$, is not an accumulation point. But as implied by the illustration, $l_8$ actually is an accumulation point of $(x_n)$.

Remarks:

1. By combining Propositions 13.47 and 13.48, we have observe that a second-countable Hausdorff space $E$ is compact iff every sequence $(x_n)$ has a convergent subsequence $(x_{n_k})$. In other words, we say a second-countable Hausdorff space $E$ is compact iff it is **sequentially compact**.

2. It should be noted that the proof showing that if $E$ is compact, then every sequence has some accumulation point, holds for any arbitrary compact space (the proof does not use a countable basis for the topology). The converse also holds for metric spaces. We will prove this converse since it is a major property of metric spaces.

Given a metric space in which every sequence has some accumulation point, we first prove the existence of a **Lebesgue number**.
Lemma 13.49. Given a metric space, $E$, if every sequence, $(x_n)$, has an accumulation point, for every open cover, $(U_i)_{i \in I}$, of $E$, there is some $\delta > 0$ (a Lebesgue number for $(U_i)_{i \in I}$) such that, for every open ball, $B_0(a, \epsilon)$, of radius $\epsilon \leq \delta$, there is some open subset, $U_i$, such that $B_0(a, \epsilon) \subseteq U_i$. See Figure 13.41

Figure 13.41: The space $E$ the closed triangular region of $\mathbb{R}^2$. It’s open cover is $(U_i)_{i=1}^8$. The Lebesgue number is the radius of the small orange balls labelled 1 through 14. Each open ball of this radius entirely contained within at least one $U_i$. For example, Ball 2 is contained in both $U_1$ and $U_2$.

Proof. If there was no $\delta$ with the above property, then, for every natural number, $n$, there would be some open ball, $B_0(a_n, 1/n)$, which is not contained in any open set, $U_i$, of the open cover, $(U_i)_{i \in I}$. However, the sequence, $(a_n)$, has some accumulation point, $a$, and since $(U_i)_{i \in I}$ is an open cover of $E$, there is some $U_i$ such that $a \in U_i$. Since $U_i$ is open, there is some open ball of center $a$ and radius $\epsilon$ contained in $U_i$. Now, since $a$ is an accumulation point of the sequence, $(a_n)$, every open set containing $a$ contains $a_n$ for infinitely many $n$ and thus, there is some $n$ large enough so that

$$1/n \leq \epsilon/2 \quad \text{and} \quad a_n \in B_0(a, \epsilon/2),$$

which implies that

$$B_0(a_n, 1/n) \subseteq B_0(a, \epsilon) \subseteq U_i,$$

a contradiction. \qed

By a previous remark, since the proof of Proposition 13.48 implies that in a compact topological space, every sequence has some accumulation point, by Lemma 13.49, in a compact metric space, every open cover has a Lebesgue number. This fact can be used to prove another important property of compact metric spaces, the uniform continuity theorem.
Definition 13.37. Given two metric spaces, \((E, d_E)\) and \((F, d_F)\), a function, \(f : E \rightarrow F\), is uniformly continuous if for every \(\epsilon > 0\), there is some \(\eta > 0\), such that, for all \(a, b \in E\),

if \(d_E(a, b) \leq \eta\) then \(d_F(f(a), f(b)) \leq \epsilon\).

See Figures 13.42 and 13.43.

![Figure 13.42: The real valued function \(f(x) = \sqrt{x}\) is uniformly continuous over \((0, \infty)\). Fix \(\epsilon\). If the \(x\) values lie within the rose colored \(\eta\) strip, the \(y\) values always lie within the peach \(\epsilon\) strip.](image)

As we saw earlier, the metric on a metric space is uniformly continuous, and the norm on a normed metric space is uniformly continuous.

The uniform continuity theorem can be stated as follows:

**Theorem 13.50.** Given two metric spaces, \((E, d_E)\) and \((F, d_F)\), if \(E\) is compact and if \(f : E \rightarrow F\) is a continuous function, then \(f\) is uniformly continuous.

**Proof.** Consider any \(\epsilon > 0\) and let \((B_0(y, \epsilon/2))_{y \in F}\) be the open cover of \(F\) consisting of open balls of radius \(\epsilon/2\). Since \(f\) is continuous, the family,

\[
(f^{-1}(B_0(y, \epsilon/2)))_{y \in F},
\]

is an open cover of \(E\). Since \(E\) is compact, by Lemma 13.49, there is a Lebesgue number, \(\delta\), such that for every open ball, \(B_0(a, \eta)\), of radius \(\eta \leq \delta\), then \(B_0(a, \eta) \subseteq f^{-1}(B_0(y, \epsilon/2))\), for some \(y \in F\). In particular, for any \(a, b \in E\) such that \(d_E(a, b) \leq \eta = \delta/2\), we have \(a, b \in B_0(a, \delta)\) and thus, \(a, b \in f^{-1}(B_0(y, \epsilon/2))\), which implies that \(f(a), f(b) \in B_0(y, \epsilon/2)\). But then, \(d_F(f(a), f(b)) \leq \epsilon\), as desired. \(\square\)

We now prove another lemma needed to obtain the characterization of compactness in metric spaces in terms of accumulation points.
Figure 13.43: The real valued function $f(x) = 1/x$ is not uniformly continuous over $(0, \infty)$. Fix $\epsilon$. In order for the $y$ values to lie within the peach epsilon strip, the widths of the eta strips decrease as $x \to 0$.

**Lemma 13.51.** Given a metric space, $E$, if every sequence, $(x_n)$, has an accumulation point, then for every $\epsilon > 0$, there is a finite open cover, $B_0(a_0, \epsilon) \cup \cdots \cup B_0(a_n, \epsilon)$, of $E$ by open balls of radius $\epsilon$.

**Proof.** Let $a_0$ be any point in $E$. If $B_0(a_0, \epsilon) = E$, then the lemma is proved. Otherwise, assume that a sequence, $(a_0, a_1, \ldots, a_n)$, has been defined, such that $B_0(a_0, \epsilon) \cup \cdots \cup B_0(a_n, \epsilon)$ does not cover $E$. Then, there is some $a_{n+1}$ not in $B_0(a_0, \epsilon) \cup \cdots \cup B_0(a_n, \epsilon)$ and either

$$B_0(a_0, \epsilon) \cup \cdots \cup B_0(a_{n+1}, \epsilon) = E,$$

in which case the lemma is proved, or we obtain a sequence, $(a_0, a_1, \ldots, a_{n+1})$, such that $B_0(a_0, \epsilon) \cup \cdots \cup B_0(a_{n+1}, \epsilon)$ does not cover $E$. If this process goes on forever, we obtain an infinite sequence, $(a_n)$, such that $d(a_m, a_n) > \epsilon$ for all $m \neq n$. Since every sequence in $E$ has some accumulation point, the sequence, $(a_n)$, has some accumulation point, $a$. Then, for infinitely many $n$, we must have $d(a_n, a) \leq \epsilon/3$ and thus, for at least two distinct natural numbers, $p, q$, we must have $d(a_p, a) \leq \epsilon/3$ and $d(a_q, a) \leq \epsilon/3$, which implies $d(a_p, a_q) \leq d(a_p, a) + d(a_q, a) \leq 2\epsilon/3$, contradicting the fact that $d(a_m, a_n) > \epsilon$ for all $m \neq n$. See Figure 13.44. Thus, there must be some $n$ such that

$$B_0(a_0, \epsilon) \cup \cdots \cup B_0(a_n, \epsilon) = E.$$

**Definition 13.38.** A metric space $E$ is said to be precompact (or totally bounded) if for every $\epsilon > 0$, there is a finite open cover, $B_0(a_0, \epsilon) \cup \cdots \cup B_0(a_n, \epsilon)$, of $E$ by open balls of radius $\epsilon$.

We now obtain the Weierstrass–Bolzano property.
Figure 13.44: Let $E$ be the peach region of $\mathbb{R}^2$. If $E$ is not covered by a finite collection of orange balls with radius $\epsilon$, the points of the sequence $(a_n)$ are separated by a distance of at least $\epsilon$. This contradicts the fact that $a$ is the accumulation point of $a$, as evidenced by the enlargement of the plum disk in Figure (ii).

**Theorem 13.52.** A metric space, $E$, is compact iff every sequence, $(x_n)$, has an accumulation point.

*Proof.* We already observed that the proof of Proposition 13.48 shows that for any compact space (not necessarily metric), every sequence, $(x_n)$, has an accumulation point. Conversely, let $E$ be a metric space, and assume that every sequence, $(x_n)$, has an accumulation point. Given any open cover, $(U_i)_{i \in I}$ for $E$, we must find a finite open subcover of $E$. By Lemma 13.49, there is some $\delta > 0$ (a Lebesgue number for $(U_i)_{i \in I}$) such that, for every open ball, $B_0(a, \epsilon)$, of radius $\epsilon \leq \delta$, there is some open subset, $U_j$, such that $B_0(a, \epsilon) \subseteq U_j$. By Lemma 13.51, for every $\delta > 0$, there is a finite open cover, $B_0(a_0, \delta) \cup \cdots \cup B_0(a_n, \delta)$, of $E$ by open balls of radius $\delta$. But from the previous statement, every open ball, $B_0(a_i, \delta)$, is contained in some open set, $U_{j_i}$, and thus, $\{U_{j_1}, \ldots, U_{j_n}\}$ is an open cover of $E$. \qed

### 13.8 Complete Metric Spaces and Compactness

Another very useful characterization of compact metric spaces is obtained in terms of Cauchy sequences. Such a characterization is quite useful in fractal geometry (and elsewhere). First
CHAPTER 13. TOPOLOGY

recall the definition of a Cauchy sequence and of a complete metric space.

Definition 13.39. Given a metric space, \((E, d)\), a sequence, \((x_n)_{n \in \mathbb{N}}\), in \(E\) is a Cauchy sequence if the following condition holds: for every \(\epsilon > 0\), there is some \(p \geq 0\), such that, for all \(m, n \geq p\), then \(d(x_m, x_n) \leq \epsilon\).

If every Cauchy sequence in \((E, d)\) converges we say that \((E, d)\) is a complete metric space.

First let us show the following proposition:

Proposition 13.53. Given a metric space, \(E\), if a Cauchy sequence, \((x_n)\), has some accumulation point, \(a\), then \(a\) is the limit of the sequence, \((x_n)\).

Proof. Since \((x_n)\) is a Cauchy sequence, for every \(\epsilon > 0\), there is some \(p \geq 0\), such that, for all \(m, n \geq p\), then \(d(x_m, x_n) \leq \epsilon/2\). Since \(a\) is an accumulation point for \((x_n)\), for infinitely many \(n\), we have \(d(x_n, a) \leq \epsilon/2\), and thus, for at least some \(n \geq p\), we have \(d(x_n, a) \leq \epsilon/2\). Then, for all \(m \geq p\),

\[
    d(x_m, a) \leq d(x_m, x_n) + d(x_n, a) \leq \epsilon,
\]

which shows that \(a\) is the limit of the sequence \((x_n)\). \(\square\)

We can now prove the following theorem.

Theorem 13.54. A metric space, \(E\), is compact iff it is precompact and complete.

Proof. Let \(E\) be compact. For every \(\epsilon > 0\), the family of all open balls of radius \(\epsilon\) is an open cover for \(E\) and since \(E\) is compact, there is a finite subcover, \(B_0(a_0, \epsilon) \cup \cdots \cup B_0(a_n, \epsilon)\), of \(E\) by open balls of radius \(\epsilon\). Thus \(E\) is precompact. Since \(E\) is compact, by Theorem 13.52, every sequence, \((x_n)\), has some accumulation point. Thus every Cauchy sequence, \((x_n)\), has some accumulation point, \(a\), and, by Proposition 13.53, \(a\) is the limit of \((x_n)\). Thus, \(E\) is complete.

Now assume that \(E\) is precompact and complete. We prove that every sequence, \((x_n)\), has an accumulation point. By the other direction of Theorem 13.52, this shows that \(E\) is compact. Given any sequence, \((x_n)\), we construct a Cauchy subsequence, \((y_n)\), of \((x_n)\) as follows: Since \(E\) is precompact, letting \(\epsilon = 1\), there exists a finite cover, \(U_1\), of \(E\) by open balls of radius 1. Thus some open ball, \(B^0_0\), in the cover, \(U_1\), contains infinitely many elements from the sequence \((x_n)\). Let \(y_0\) be any element of \((x_n)\) in \(B^0_0\). By induction, assume that a sequence of open balls, \((B^i_0)_{1 \leq i \leq m}\), has been defined, such that every ball, \(B^i_0\), has radius \(\frac{1}{2^i}\), contains infinitely many elements from the sequence \((x_n)\) and contains some \(y_i\) from \((x_n)\) such that

\[
    d(y_i, y_{i+1}) \leq \frac{1}{2^i},
\]

for all \(i, 0 \leq i \leq m - 1\). See Figure 13.45. Then letting \(\epsilon = \frac{1}{2^m+\epsilon}\), because \(E\) is precompact, there is some finite cover, \(U_{m+1}\), of \(E\) by open balls of radius \(\epsilon\) and thus, of the open ball \(B^m_0\).
Thus, some open ball, \( B_{\delta}^{m+1} \), in the cover, \( \mathcal{U}_{m+1} \), contains infinitely many elements from the sequence, \((x_n)\), and we let \( y_{m+1} \) be any element of \((x_n)\) in \( B_{\delta}^{m+1} \). Thus, we have defined by induction a sequence, \((y_n)\), which is a subsequence of, \((x_n)\), and such that
\[
d(y_i, y_{i+1}) \leq \frac{1}{2^i},
\]
for all \( i \). However, for all \( m, n \geq 1 \), we have
\[
d(y_m, y_n) \leq d(y_m, y_{m+1}) + \cdots + d(y_{n-1}, y_n) \leq \sum_{i=m}^{n} \frac{1}{2^i} \leq \frac{1}{2^{m-1}},
\]
and thus, \((y_n)\) is a Cauchy sequence. Since \( E \) is complete, the sequence, \((y_n)\), has a limit, and since it is a subsequence of \((x_n)\), the sequence, \((x_n)\), has some accumulation point.

Another useful property of a complete metric space is that a subset is closed iff it is complete. This is shown in the following two propositions.

**Proposition 13.55.** Let \((E, d)\) be a metric space, and let \( A \) be a subset of \( E \). If \( A \) is complete (which means that every Cauchy sequence of elements in \( A \) converges to some point of \( A \)), then \( A \) is closed in \( E \).
Proof. Assume \( x \in \overline{A} \). By Proposition 13.13, there is some sequence \((a_n)\) of points \( a_n \in A \) which converges to \( x \). Consequently \((a_n)\) is a Cauchy sequence in \( E \), and thus a Cauchy sequence in \( A \) (since \( a_n \in A \) for all \( n \)). Since \( A \) is complete, the sequence \((a_n)\) has a limit \( a \in A \), but since \( E \) is a metric space it is Hausdorff, so \( a = x \), which shows that \( x \in A \); that is, \( A \) is closed.

Proposition 13.56. Let \((E,d)\) be a metric space, and let \( A \) be a subset of \( E \). If \( E \) is complete and if \( A \) is closed in \( E \), then \( A \) is complete.

Proof. Let \( (a_n) \) be a Cauchy sequence in \( A \). The sequence \((a_n)\) is also a Cauchy sequence in \( E \), and since \( E \) is complete, it has a limit \( x \in E \). But \( a_n \in A \) for all \( n \), so by Proposition 13.13 we must have \( x \in \overline{A} \). Since \( A \) is closed, actually \( x \in A \), which proves that \( A \) is complete. \( \square \)

An arbitrary metric space \((E,d)\) is not necessarily complete, but there is a construction of a metric space \((\hat{E},\hat{d})\) such that \( \hat{E} \) is complete, and there is a continuous (injective) distance-preserving map \( \varphi: E \to \hat{E} \) such that \( \varphi(E) \) is dense in \( \hat{E} \). This is a generalization of the construction of the set \( \mathbb{R} \) of real numbers from the set \( \mathbb{Q} \) of rational numbers in terms of Cauchy sequences. This construction can be immediately adapted to a normed vector space \((E,\|\|)\) to embed \((E,\|\|)\) into a complete normed vector space \((\hat{E},\|\|_{\hat{E}})\) (a Banach space). This construction is used heavily in integration theory, where \( E \) is a set of functions.

### 13.9 Completion of a Metric Space

In order to prove a kind of uniqueness result for the completion \((\hat{E},\hat{d})\) of a metric space \((E,d)\), we need the following result about extending a uniformly continuous function.

Recall that \( E_0 \) is dense in \( E \) iff \( \overline{E_0} = E \). Since \( E \) is a metric space, by Proposition 13.13, this means that for every \( x \in E \), there is some sequence \((x_n)\) converging to \( x \), with \( x_n \in E_0 \).

Theorem 13.57. Let \( E \) and \( F \) be two metric spaces, let \( E_0 \) be a dense subspace of \( E \), and let \( f_0: E_0 \to F \) be a continuous function. If \( f_0 \) is uniformly continuous and if \( F \) is complete, then there is a unique uniformly continuous function \( f: E \to F \) extending \( f_0 \).

Proof. We follow Schwartz’s proof; see Schwartz [38] (Chapter XI, Section 3, Theorem 1).

Step 1. We begin by constructing a function \( f: E \to F \) extending \( f_0 \). Since \( E_0 \) is dense in \( E \), for every \( x \in E \), there is some sequence \((x_n)\) converging to \( x \), with \( x_n \in E_0 \). Then the sequence \((x_n)\) is a Cauchy sequence in \( E \). We claim that \((f_0(x_n))\) is a Cauchy sequence in \( F \).

Proof of the claim. For every \( \epsilon > 0 \), since \( f_0 \) is uniformly continuous, there is some \( \eta > 0 \) such that for all \((y,z) \in E_0\), if \( d(y,z) \leq \eta \), then \( d(f_0(y), f_0(z)) \leq \epsilon \). Since \((x_n)\) is a Cauchy sequence with \( x_n \in E_0 \), there is some integer \( p > 0 \) such that if \( m, n \geq p \), then \( d(x_m, x_n) \leq \eta \), thus \( d(f_0(x_m), f_0(x_n)) \leq \epsilon \), which proves that \((f_0(x_n))\) is a Cauchy sequence in \( F \). \( \square \)
Since $F$ is complete and $(f_0(x_n))$ is a Cauchy sequence in $F$, the sequence $(f_0(x_n))$ converges to some element of $F$; denote this element by $f(x)$.

**Step 2.** Let us now show that $f(x)$ does not depend on the sequence $(x_n)$ converging to $x$. Suppose that $(x_n')$ and $(x_n'')$ are two sequences of elements in $E_0$ converging to $x$. Then the mixed sequence

$$x_0', x_0'', x_1', x_1'', \ldots, x_n', x_n'', \ldots,$$

also converges to $x$. It follows that the sequence

$$f_0(x_0'), f_0(x_0''), f_0(x_1'), f_0(x_1''), \ldots, f_0(x_n'), f_0(x_n''), \ldots,$$

is a Cauchy sequence in $F$, and since $F$ is complete, it converges to some element of $F$, which implies that the sequences $(f_0(x_n'))$ and $(f_0(x_n''))$ converge to the same limit.

As a summary, we have defined a function $f : E \to F$ by

$$f(x) = \lim_{n \to \infty} f_0(x_n),$$

for any sequence $(x_n)$ converging to $x$, with $x_n \in E_0$.

**Step 3.** The function $f$ extends $f_0$. Since every element $x \in E_0$ is the limit of the constant sequence $(x_n)$ with $x_n = x$ for all $n \geq 0$, by definition $f(x)$ is the limit of the sequence $(f_0(x_n))$, which is the constant sequence with value $f_0(x)$, so $f(x) = f_0(x)$; that is, $f$ extends $f_0$.

**Step 4.** We now prove that $f$ is uniformly continuous. Since $f_0$ is uniformly continuous, for every $\epsilon > 0$, there is some $\eta > 0$ such that if $a, b \in E_0$ and $d(a, b) \leq \eta$, then $d(f_0(a), f_0(b)) \leq \epsilon$. Consider any two points $x, y \in E$ such that $d(x, y) \leq \eta/2$. We claim that $d(f(x), f(y)) \leq \epsilon$, which shows that $f$ is uniformly continuous.

Let $(x_n)$ be a sequence of points in $E_0$ converging to $x$, and let $(y_n)$ be a sequence of points in $E_0$ converging to $y$. By the triangle inequality,

$$d(x_n, y_n) \leq d(x_n, x) + d(x, y) + d(y, y_n) = d(x, y) + d(x_n, x) + d(y_n, y),$$

and since $(x_n)$ converges to $x$ and $(y_n)$ converges to $y$, there is some integer $p > 0$ such that for all $n \geq p$, we have $d(x_n, x) \leq \eta/4$ and $d(y_n, y) \leq \eta/4$, and thus

$$d(x_n, y_n) \leq d(x, y) + \frac{\eta}{2}.$$

Since we assumed that $d(x, y) \leq \eta/2$, we get $d(x_n, y_n) \leq \eta$ for all $n \geq p$, and by uniform continuity of $f_0$, we get

$$d(f_0(x_n), f_0(y_n)) \leq \epsilon$$

for all $n \geq p$. Since the distance function on $F$ is also continuous, and since $(f_0(x_n))$ converges to $f(x)$ and $(f_0(y_n))$ converges to $f(y)$, we deduce that the sequence $(d(f_0(x_n), f_0(y_n)))$ converges to $d(f(x), f(y))$. This implies that $d(f(x), f(y)) \leq \epsilon$, as desired.
Step 5. It remains to prove that $f$ is unique. Since $E_0$ is dense in $E$, for every $x \in E$, there is some sequence $(x_n)$ converging to $x$, with $x_n \in E_0$. Since $f$ extends $f_0$ and since $f$ is continuous, we get

$$f(x) = \lim_{n \to \infty} f_0(x_n),$$

which only depends on $f_0$ and $x$, and shows that $f$ is unique. \qed

Remark: It can be shown that the theorem no longer holds if we either omit the hypothesis that $F$ is complete or omit that $f_0$ is uniformly continuous.

For example, if $E_0 \neq E$ and if we let $F = E_0$ and $f_0$ be the identity function, it is easy to see that $f_0$ cannot be extended to a continuous function from $E$ to $E_0$ (for any $x \in E - E_0$, any continuous extension $f$ of $f_0$ would satisfy $f(x) = x$, which is absurd since $x \notin E_0$).

If $f_0$ is continuous but not uniformly continuous, a counter-example can be given by using $E = \mathbb{R} = \mathbb{R} \cup \{\infty\}$ made into a metric space, $E_0 = \mathbb{R}$, $F = \mathbb{R}$, and $f_0$ the identity function; for details, see Schwartz [38] (Chapter XI, Section 3, page 134).

Definition 13.40. If $(E,d_E)$ and $(F,d_F)$ are two metric spaces, then a function $f : E \to F$ is distance-preserving, or an isometry, if

$$d_F(f(x), f(y)) = d_E(x, y), \quad \text{for all for all } x, y \in E.$$

Observe that an isometry must be injective, because if $f(x) = f(y)$, then $d_F(f(x), f(y)) = 0$, and since $d_F(f(x), f(y)) = d_E(x, y)$, we get $d_E(x, y) = 0$, but $d_E(x, y) = 0$ implies that $x = y$. Also, an isometry is uniformly continuous (since we can pick $\eta = \epsilon$ to satisfy the condition of uniform continuity). However, an isometry is not necessarily surjective.

We now give a construction of the completion of a metric space. This construction is just a generalization of the classical construction of $\mathbb{R}$ from $\mathbb{Q}$ using Cauchy sequences.

Theorem 13.58. Let $(E,d)$ be any metric space. There is a complete metric space $(\widehat{E}, \widehat{d})$ called a completion of $(E,d)$, and a distance-preserving (uniformly continuous) map $\varphi : E \to \widehat{E}$ such that $\varphi(E)$ is dense in $\widehat{E}$, and the following extension property holds: for every complete metric space $F$ and for every uniformly continuous function $f : E \to F$, there is a unique uniformly continuous function $\widehat{f} : \widehat{E} \to F$ such that

$$f = \widehat{f} \circ \varphi,$$

as illustrated in the following diagram.

$$\begin{array}{ccc}
E & \xrightarrow{\varphi} & \widehat{E} \\
\downarrow f & & \downarrow \widehat{f} \\
& F.&
\end{array}$$

As a consequence, for any two completions $(\widehat{E}_1, \widehat{d}_1)$ and $(\widehat{E}_2, \widehat{d}_2)$ of $(E,d)$, there is a unique bijective isometry between $(\widehat{E}_1, \widehat{d}_1)$ and $(\widehat{E}_2, \widehat{d}_2)$. 
13.9. COMPLETION OF A METRIC SPACE

Proof. Consider the set $E$ of all Cauchy sequences $(x_n)$ in $E$, and define the relation $\sim$ on $E$ as follows:

$$(x_n) \sim (y_n) \text{ iff } \lim_{n \to \infty} d(x_n, y_n) = 0.$$ 

It is easy to check that $\sim$ is an equivalence relation on $E$, and let $\hat{E} = E/\sim$ be the quotient set, that is, the set of equivalence classes modulo $\sim$. Our goal is to show that we can endow $\hat{E}$ with a distance that makes it into a complete metric space satisfying the conditions of the theorem. We proceed in several steps.

**Step 1.** First, let us construct the function $\varphi : E \to \hat{E}$. For every $a \in E$, we have the constant sequence $(a_n)$ such that $a_n = a$ for all $n \geq 0$, which is obviously a Cauchy sequence. Let $\varphi(a) \in \hat{E}$ be the equivalence class $[(a_n)]$ of the constant sequence $(a_n)$ with $a_n = a$ for all $n$. By definition of $\sim$, the equivalence class $\varphi(a)$ is also the equivalence class of all sequences converging to $a$. The map $a \mapsto \varphi(a)$ is injective because a metric space is Hausdorff, so if $a \neq b$, then a sequence converging to $a$ does not converge to $b$. After having defined a distance on $\hat{E}$, we will check that $\varphi$ is an isometry.

**Step 2.** Let us now define a distance on $\hat{E}$. Let $\alpha = [(a_n)]$ and $\beta = [(b_n)]$ be two equivalence classes of Cauchy sequences in $E$. The triangle inequality implies that

$$d(a_m, b_m) \leq d(a_m, a_n) + d(a_n, b_n) + d(b_n, b_m) = d(a_n, b_n) + d(a_m, a_n) + d(b_m, b_n)$$

and

$$d(a_n, b_n) \leq d(a_n, a_m) + d(a_m, b_m) + d(b_m, b_n) = d(a_m, b_n) + d(a_m, a_n) + d(b_m, b_n),$$

which implies that

$$|d(a_m, b_m) - d(a_n, b_n)| \leq d(a_m, a_n) + d(b_m, b_n).$$

Since $(a_n)$ and $(b_n)$ are Cauchy sequences, it follows that $(d(a_n, b_n))$ is a Cauchy sequence of nonnegative reals. Since $\mathbb{R}$ is complete, the sequence $(d(a_n, b_n))$ has a limit, which we denote by $\tilde{d}(\alpha, \beta)$; that is, we set

$$\tilde{d}(\alpha, \beta) = \lim_{n \to \infty} d(a_n, b_n), \quad \alpha = [(a_n)], \quad \beta = [(b_n)].$$

**Step 3.** Let us check that $\tilde{d}(\alpha, \beta)$ does not depend on the Cauchy sequences $(a_n)$ and $(b_n)$ chosen in the equivalence classes $\alpha$ and $\beta$.

If $(a_n) \sim (a'_n)$ and $(b_n) \sim (b'_n)$, then $\lim_{n \to \infty} d(a_n, a'_n) = 0$ and $\lim_{n \to \infty} d(b_n, b'_n) = 0$, and since

$$d(a'_n, b'_n) \leq d(a'_n, a_n) + d(a_n, b_n) + d(b_n, b'_n) = d(a_n, b_n) + d(a_n, a'_n) + d(b_n, b'_n)$$

and

$$d(a_n, b_n) \leq d(a_n, a'_n) + d(a'_n, b'_n) + d(b'_n, b_n) = d(a'_n, b'_n) + d(a_n, a'_n) + d(b_n, b'_n)$$

we have $\tilde{d}(\alpha, \beta) = \lim_{n \to \infty} d(a_n, b_n)$. Therefore $\tilde{d}(\alpha, \beta)$ does not depend on the Cauchy sequences $(a_n)$ and $(b_n)$ chosen in the equivalence classes $\alpha$ and $\beta$. For every $\alpha, \beta \in \hat{E}$, the function $d : \hat{E} \times \hat{E} \to \mathbb{R}$ defined by $d(\alpha, \beta) = \tilde{d}(\alpha, \beta)$ is well defined, and it is easy to check that $d$ is a distance on $\hat{E}$ that makes it into a complete metric space.
we have
\[ |d(a_n, b_n) - d(a'_n, b'_n)| \leq d(a_n, a'_n) + d(b_n, b'_n), \]
so we have \( \lim_{n \to \infty} d(a'_n, b'_n) = \lim_{n \to \infty} d(a_n, b_n) = \hat{d}(\alpha, \beta). \) Therefore, \( \hat{d}(\alpha, \beta) \) is indeed well defined.

\textbf{Step 4.} Let us check that \( \varphi \) is indeed an isometry.

Given any two elements \( \varphi(a) \) and \( \varphi(b) \) in \( \hat{E} \), since they are the equivalence classes of the constant sequences \( (a_n) \) and \( (b_n) \) such that \( a_n = a \) and \( b_n = b \) for all \( n \), the constant sequence \( (d(a_n, b_n)) \) with \( d(a_n, b_n) = d(a, b) \) for all \( n \) converges to \( d(a, b) \), so by definition \( \hat{d}(\varphi(a), \varphi(b)) = \lim_{n \to \infty} d(a_n, b_n) = d(a, b) \), which shows that \( \varphi \) is an isometry.

\textbf{Step 5.} Let us verify that \( \hat{d} \) is a metric on \( \hat{E} \). By definition it is obvious that \( \hat{d}(\alpha, \beta) = \hat{d}(\beta, \alpha) \). If \( \alpha \) and \( \beta \) are two distinct equivalence classes, then for any Cauchy sequence \( (a_n) \) in the equivalence class \( \alpha \) and for any Cauchy sequence \( (b_n) \) in the equivalence class \( \beta \), the sequences \( (a_n) \) and \( (b_n) \) are inequivalent, which means that \( \lim_{n \to \infty} d(a_n, b_n) \neq 0 \), that is, \( \hat{d}(\alpha, \beta) \neq 0 \). Obviously, \( \hat{d}(\alpha, \alpha) = 0 \).

For any equivalence classes \( \alpha = [(a_n)], \beta = [(b_n)], \) and \( \gamma = [(c_n)] \), we have the triangle inequality
\[ d(a_n, c_n) \leq d(a_n, b_n) + d(b_n, c_n), \]
so by continuity of the distance function, by passing to the limit, we obtain
\[ \hat{d}(\alpha, \gamma) \leq \hat{d}(\alpha, \beta) + \hat{d}(\beta, \gamma), \]
which is the triangle inequality for \( \hat{d} \). Therefore, \( \hat{d} \) is a distance on \( \hat{E} \).

\textbf{Step 6.} Let us prove that \( \varphi(E) \) is dense in \( \hat{E} \). For any \( \alpha = [(a_n)] \), let \( (x_n) \) be the constant sequence such that \( x_k = a_n \) for all \( k \geq 0 \), so that \( \varphi(a_n) = [(x_n)] \). Then we have
\[ \hat{d}(\alpha, \varphi(a_n)) = \lim_{m \to \infty} d(a_m, a_n) \leq \sup_{p,q \geq n} d(a_p, a_q). \]
Since \( (a_n) \) is a Cauchy sequence, \( \sup_{p,q \geq n} d(a_p, a_q) \) tends to 0 as \( n \) goes to infinity, so
\[ \lim_{n \to \infty} d(\alpha, \varphi(a_n)) = 0, \]
which means that the sequence \( (\varphi(a_n)) \) converge to \( \alpha \), and \( \varphi(E) \) is indeed dense in \( \hat{E} \).

\textbf{Step 7.} Finally, let us prove that the metric space \( \hat{E} \) is complete.

Let \( (\alpha_n) \) be a Cauchy sequence in \( \hat{E} \). Since \( \varphi(E) \) is dense in \( \hat{E} \), for every \( n > 0 \), there some \( a_n \in E \) such that
\[ \hat{d}(\alpha_n, \varphi(a_n)) \leq \frac{1}{n}. \]
Since
\[ \hat{d}(\varphi(a_m), \varphi(a_n)) \leq \hat{d}(\varphi(a_m), \alpha_m) + \hat{d}(\alpha_m, \alpha_n) + \hat{d}(\alpha_n, \varphi(a_n)) \leq \hat{d}(\alpha_m, \alpha_n) + \frac{1}{m} + \frac{1}{n}, \]
and since \((\alpha_m)\) is a Cauchy sequence, so is \((\varphi(a_n))\), and as \(\varphi\) is an isometry, the sequence \((a_n)\) is a Cauchy sequence in \(E\). Let \(\alpha \in \hat{E}\) be the equivalence class of \((a_n)\). Since
\[
\hat{d}(\alpha, \varphi(a_n)) = \lim_{m \to \infty} d(a_m, a_n)
\]
and \((a_n)\) is a Cauchy sequence, we deduce that the sequence \((\varphi(a_n))\) converges to \(\alpha\), and since \(d(\alpha_n, \varphi(a_n)) \leq 1/n\) for all \(n > 0\), the sequence \((\alpha_n)\) also converges to \(\alpha\).

**Step 8.** Let us prove the extension property. Let \(F\) be any complete metric space and let \(f: E \to F\) be any uniformly continuous function. The function \(\varphi: E \to \hat{E}\) is an isometry and a bijection between \(E\) and its image \(\varphi(E)\), so its inverse \(\varphi^{-1}: \varphi(E) \to E\) is also an isometry, and thus is uniformly continuous. If we let \(g = f \circ \varphi^{-1}\), then \(g: \varphi(E) \to F\) is a uniformly continuous function, and \(\varphi(E)\) is dense in \(\hat{E}\), so by Theorem 13.57 there is a unique uniformly continuous function \(\hat{f}: \hat{E} \to F\) extending \(g = f \circ \varphi^{-1}\); see the diagram below:

\[
\begin{array}{ccc}
E & \xrightarrow{\varphi^{-1}} & \varphi(E) \subseteq \hat{E} \\
\downarrow{f} & & \downarrow{g} \\
F & \xleftarrow{\hat{f}} & \\
\end{array}
\]

This means that
\[
\hat{f}|\varphi(E) = f \circ \varphi^{-1},
\]
which implies that
\[
(\hat{f}|\varphi(E)) \circ \varphi = f,
\]
that is, \(f = \hat{f} \circ \varphi\), as illustrated in the diagram below:

\[
\begin{array}{ccc}
E & \xrightarrow{\varphi} & \hat{E} \\
\downarrow{f} & & \downarrow{\hat{f}} \\
F & & \\
\end{array}
\]

If \(h: \hat{E} \to F\) is any other uniformly continuous function such that \(f = h \circ \varphi\), then \(g = f \circ \varphi^{-1} = h|\varphi(E)\), so \(h\) is a uniformly continuous function extending \(g\), and by Theorem 13.57, we have have \(h = \hat{f}\), so \(\hat{f}\) is indeed unique.

**Step 9.** Uniqueness of the completion \((\hat{E}, \hat{d})\) up to a bijective isometry.

Let \((\hat{E}_1, \hat{d}_1)\) and \((\hat{E}_2, \hat{d}_2)\) be any two completions of \((E, d)\). Then we have two uniformly continuous isometries \(\varphi_1: E \to \hat{E}_1\) and \(\varphi_2: E \to \hat{E}_2\), so by the unique extension property, there exist unique uniformly continuous maps \(\hat{\varphi}_2: \hat{E}_1 \to \hat{E}_2\) and \(\hat{\varphi}_1: \hat{E}_2 \to \hat{E}_1\) such that the following diagrams commute:

\[
\begin{array}{ccc}
E & \xrightarrow{\varphi_1} & \hat{E}_1 \\
\downarrow{\varphi_2} & & \downarrow{\hat{\varphi}_2} \\
\hat{E}_2 & & \\
\end{array}
\quad
\begin{array}{ccc}
E & \xrightarrow{\varphi_2} & \hat{E}_2 \\
\downarrow{\varphi_1} & & \downarrow{\hat{\varphi}_1} \\
\hat{E}_1 & & \\
\end{array}
\]
Consequently we have the following commutative diagrams:

\[ \begin{array}{ccc}
E & \xrightarrow{\varphi_1} & \hat{E}_1 \\
\downarrow & & \downarrow \\
\hat{E}_2 & \xrightarrow{\varphi_2} & \hat{E}_2 \\
\end{array} \quad \begin{array}{ccc}
E & \xrightarrow{\varphi_2} & \hat{E}_2 \\
\downarrow & & \downarrow \\
\hat{E}_1 & \xrightarrow{\varphi_1} & \hat{E}_1 \\
\end{array} \]

However, id\(_{\hat{E}_1}\) and id\(_{\hat{E}_2}\) are uniformly continuous functions making the following diagrams commute

\[ \begin{array}{ccc}
E & \xrightarrow{\varphi_1} & \hat{E}_1 \\
\downarrow & & \downarrow \\
\hat{E}_1 & \xrightarrow{id_{\hat{E}_1}} & \hat{E}_1 \\
\end{array} \quad \begin{array}{ccc}
E & \xrightarrow{\varphi_2} & \hat{E}_2 \\
\downarrow & & \downarrow \\
\hat{E}_2 & \xrightarrow{id_{\hat{E}_2}} & \hat{E}_2 \\
\end{array} \]

so by the uniqueness of extensions we must have

\[ \varphi_1 \circ \varphi_2 = id_{\hat{E}_1} \quad \text{and} \quad \varphi_2 \circ \varphi_1 = id_{\hat{E}_2}. \]

This proves that \(\varphi_1\) and \(\varphi_2\) are mutual inverses. Now, since \(\varphi_2 = \varphi_2 \circ \varphi_1\), we have

\[ \varphi_2|\varphi_1(E) = \varphi_2 \circ \varphi_1^{-1}, \]

and since \(\varphi_1^{-1}\) and \(\varphi_2\) are isometries, so is \(\varphi_2|\varphi_1(E)\). But we saw earlier that \(\varphi_2\) is the uniform continuous extension of \(\varphi_2|\varphi_1(E)\) and \(\varphi_1(E)\) is dense in \(\hat{E}_1\), so for any two elements \(\alpha, \beta \in \hat{E}_1\), if \((a_n)\) and \((b_n)\) are sequences in \(\varphi_1(E)\) converging to \(\alpha\) and \(\beta\), we have

\[ \widehat{d}_2((\varphi_2|\varphi_1(E))(a_n), (\varphi_2|\varphi_1(E))(b_n)) = \widehat{d}_1(a_n, b_n), \]

and by passing to the limit we get

\[ \widehat{d}_2(\varphi_2(\alpha), \varphi_2(\beta)) = \widehat{d}_1(\alpha, \beta), \]

which shows that \(\varphi_2\) is an isometry (similarly, \(\varphi_1\) is an isometry).

\[ \square \]

**Remarks:**

1. Except for Step 8 and Step 9, the proof of Theorem 13.58 is the proof given in Schwartz [38] (Chapter XI, Section 4, Theorem 1), and Kormogorov and Fomin [23] (Chapter 2, Section 7, Theorem 4).

2. The construction of \(\hat{E}\) relies on the completeness of \(\mathbb{R}\), and so it cannot be used to construct \(\mathbb{R}\) from \(\mathbb{Q}\). However, this construction can be modified to yield a construction of \(\mathbb{R}\) from \(\mathbb{Q}\).

We show in Section 13.12 that Theorem 13.58 yields a construction of the completion of a normed vector space.
13.10 The Contraction Mapping Theorem

If \((E, d)\) is a nonempty complete metric space, every map, \(f : E \to E\), for which there is some \(k\) such that \(0 \leq k < 1\) and

\[d(f(x), f(y)) \leq kd(x, y)\]

for all \(x, y \in E\), has the very important property that it has a unique fixed point, that is, there is a unique, \(a \in E\), such that \(f(a) = a\). A map as above is called a contraction mapping. Furthermore, the fixed point of a contraction mapping can be computed as the limit of a fast converging sequence.

The fixed point property of contraction mappings is used to show some important theorems of analysis, such as the implicit function theorem and the existence of solutions to certain differential equations. It can also be used to show the existence of fractal sets defined in terms of iterated function systems. Since the proof is quite simple, we prove the fixed point property of contraction mappings. First, observe that a contraction mapping is (uniformly) continuous.

**Proposition 13.59.** If \((E, d)\) is a nonempty complete metric space, every contraction mapping, \(f : E \to E\), has a unique fixed point. Furthermore, for every \(x_0 \in E\), defining the sequence, \((x_n)\), such that \(x_{n+1} = f(x_n)\), the sequence, \((x_n)\), converges to the unique fixed point of \(f\).

**Proof.** First we prove that \(f\) has at most one fixed point. Indeed, if \(f(a) = a\) and \(f(b) = b\), since

\[d(a, b) = d(f(a), f(b)) \leq kd(a, b)\]

and \(0 \leq k < 1\), we must have \(d(a, b) = 0\), that is, \(a = b\).

Next, we prove that \((x_n)\) is a Cauchy sequence. Observe that

\[
\begin{align*}
    d(x_2, x_1) & \leq kd(x_1, x_0), \\
    d(x_3, x_2) & \leq kd(x_2, x_1) \leq k^2d(x_1, x_0), \\
    & \vdots \\
    d(x_{n+1}, x_n) & \leq kd(x_n, x_{n-1}) \leq \cdots \leq k^n d(x_1, x_0).
\end{align*}
\]

Thus, we have

\[
\begin{align*}
    d(x_{n+p}, x_n) & \leq d(x_{n+p}, x_{n+p-1}) + d(x_{n+p-1}, x_{n+p-2}) + \cdots + d(x_{n+1}, x_n) \\
    & \leq (k^{p-1} + k^{p-2} + \cdots + k + 1)kd(x_1, x_0) \\
    & \leq \frac{k^n}{1-k} d(x_1, x_0).
\end{align*}
\]
We conclude that \( d(x_{n+p}, x_n) \) converges to 0 when \( n \) goes to infinity, which shows that \( (x_n) \) is a Cauchy sequence. Since \( E \) is complete, the sequence \( (x_n) \) has a limit, \( a \). Since \( f \) is continuous, the sequence \( (f(x_n)) \) converges to \( f(a) \). But \( x_{n+1} = f(x_n) \) converges to \( a \) and so \( f(a) = a \), the unique fixed point of \( f \).

Note that no matter how the starting point \( x_0 \) of the sequence \( (x_n) \) is chosen, \( (x_n) \) converges to the unique fixed point of \( f \). Also, the convergence is fast, since
\[
d(x_n, a) \leq \frac{k^n}{1 - k} d(x_1, x_0).
\]

The Hausdorff distance between compact subsets of a metric space provides a very nice illustration of some of the theorems on complete and compact metric spaces just presented.

**Definition 13.41.** Given a metric space, \((X, d)\), for any subset, \(A \subseteq X\), for any, \(\epsilon \geq 0\), define the \(\epsilon\)-hull of \(A\) as the set
\[
V_\epsilon(A) = \{ x \in X, \exists a \in A \mid d(a, x) \leq \epsilon \}.
\]

See Figure 13.46. Given any two nonempty bounded subsets, \(A, B\) of \(X\), define \(D(A, B)\), the Hausdorff distance between \(A\) and \(B\), by
\[
D(A, B) = \inf\{ \epsilon \geq 0 \mid A \subseteq V_\epsilon(B) \text{ and } B \subseteq V_\epsilon(A) \}\.
\]

![Figure 13.46: The \(\epsilon\)-hull of a polygonal region \(A\) of \(\mathbb{R}^2\)](image)

Note that since we are considering nonempty bounded subsets, \(D(A, B)\) is well defined (i.e., not infinite). However, \(D\) is not necessarily a distance function. It is a distance function if we restrict our attention to nonempty compact subsets of \(X\) (actually, it is also a metric on closed and bounded subsets). We let \(\mathcal{K}(X)\) denote the set of all nonempty compact subsets of \(X\). The remarkable fact is that \(D\) is a distance on \(\mathcal{K}(X)\) and that if \(X\) is complete or compact, then so is \(\mathcal{K}(X)\). The following theorem is taken from Edgar [17].
Theorem 13.60. If \((X, d)\) is a metric space, then the Hausdorff distance, \(D\), on the set, \(K(X)\), of nonempty compact subsets of \(X\) is a distance. If \((X, d)\) is complete, then \((K(X), D)\) is complete and if \((X, d)\) is compact, then \((K(X), D)\) is compact.

Proof. Since (nonempty) compact sets are bounded, \(D(A, B)\) is well defined. Clearly \(D\) is symmetric. Assume that \(D(A, B) = 0\). Then for every \(\epsilon > 0\), \(A \subseteq V_\epsilon(B)\), which means that for every \(a \in A\), there is some \(b \in B\) such that \(d(a, b) \leq \epsilon\), and thus, that \(A \subseteq \overline{B}\). Since Proposition 13.26 implies that \(B\) is closed, \(\overline{B} = B\), and we have \(A \subseteq B\). Similarly, \(B \subseteq A\), and thus, \(A = B\). Clearly, if \(A = B\), we have \(D(A, B) = 0\). It remains to prove the triangle inequality. Assume that \(D(A, B) \leq \epsilon_1\) and that \(D(B, C) \leq \epsilon_2\). We must show that \(D(A, C) \leq \epsilon_1 + \epsilon_2\). This will be accomplished if we can show that \(C \subseteq V_{\epsilon_1+\epsilon_2}(A)\) and \(A \subseteq V_{\epsilon_1+\epsilon_2}(C)\). By assumption and definition of \(D\), \(B \subseteq V_{\epsilon_1}(A)\) and \(C \subseteq V_{\epsilon_2}(B)\). Then

\[ V_{\epsilon_2}(B) \subseteq V_{\epsilon_2}(V_{\epsilon_1}(A)) \]

and since a basic application of the triangle inequality implies that

\[ V_{\epsilon_2}(V_{\epsilon_1}(A)) \subseteq V_{\epsilon_1+\epsilon_2}(A), \]

we get

\[ C \subseteq V_{\epsilon_2}(B) \subseteq V_{\epsilon_1+\epsilon_2}(A). \]

See Figure 13.47.

Figure 13.47: Let \(A\) be the small pink square and \(B\) be the small purple triangle in \(\mathbb{R}^2\). The periwinkle oval \(C\) is contained in \(V_{\epsilon_1+\epsilon_2}(A)\).

Similarly, the conditions \((A, B) \leq \epsilon_1\) and \(D(B, C) \leq \epsilon_2\) imply that

\[ A \subseteq V_{\epsilon_1}(B), \quad B \subseteq V_{\epsilon_2}(C). \]
Hence

$$A \subseteq V_{\epsilon_1}(B) \subseteq V_{\epsilon_1}(V_{\epsilon_2}(C)) \subseteq V_{\epsilon_1+\epsilon_2}(C),$$

and thus the triangle inequality follows.

Next we need to prove that if \((X,d)\) is complete, then \((\mathcal{K}(X),D)\) is also complete. First we show that if \((A_n)\) is a sequence of nonempty compact sets converging to a nonempty compact set \(A\) in the Hausdorff metric, then

$$A = \{x \in X \mid \text{there is a sequence, } (x_n), \text{ with } x_n \in A_n \text{ converging to } x \}.$$ 

Indeed, if \((x_n)\) is a sequence with \(x_n \in A_n\) converging to \(x\) and \((A_n)\) converges to \(A\) then, for every \(\epsilon > 0\), there is some \(x_n\) such that \(d(x_n, x) \leq \epsilon/2\) and there is some \(a_n \in A\) such that \(d(a_n, x_n) \leq \epsilon/2\) and thus, \(d(a_n, x) \leq \epsilon\), which shows that \(x \in \overline{A}\). Since \(A\) is compact, it is closed, and \(x \in A\). See Figure 13.48.

Figure 13.48: Let \((A_n)\) be the sequence of parallelograms converging to \(A\), the large pale yellow parallelogram. Figure (ii.) expands the dashed region and shows why \(d(a_n, x) < \epsilon\).

Conversely, since \((A_n)\) converges to \(A\), for every \(x \in A\), for every \(n \geq 1\), there is some \(x_n \in A_n\) such that \(d(x_n, x) \leq 1/n\) and the sequence \((x_n)\) converges to \(x\).

Now let \((A_n)\) be a Cauchy sequence in \(\mathcal{K}(X)\). It can be proven that \((A_n)\) converges to the set

$$A = \{x \in X \mid \text{there is a sequence, } (x_n), \text{ with } x_n \in A_n \text{ converging to } x \},$$

and that \(A\) is nonempty and compact. To prove that \(A\) is compact, one proves that it is totally bounded and complete. Details are given in Edgar [17].
Finally we need to prove that if \((X, d)\) is compact, then \((\mathcal{K}(X), D)\) is compact. Since we already know that \((\mathcal{K}(X), D)\) is complete if \((X, d)\) is, it is enough to prove that \((\mathcal{K}(X), D)\) is totally bounded if \((X, d)\) is, which is not hard.

In view of Theorem 13.60 and Theorem 13.59, it is possible to define some nonempty compact subsets of \(X\) in terms of fixed points of contraction maps. This can be done in terms of iterated function systems, yielding a large class of fractals. However, we will omit this topic and instead refer the reader to Edgar [17].

Before considering differentials, we need to look at the continuity of linear maps.

## 13.11 Continuous Linear and Multilinear Maps

If \(E\) and \(F\) are normed vector spaces, we first characterize when a linear map \(f : E \to F\) is continuous.

**Proposition 13.61.** Given two normed vector spaces \(E\) and \(F\), for any linear map \(f : E \to F\), the following conditions are equivalent:

1. The function \(f\) is continuous at 0.
2. There is a constant \(k \geq 0\) such that,
   \[ \|f(u)\| \leq k, \text{ for every } u \in E \text{ such that } \|u\| \leq 1. \]
3. There is a constant \(k \geq 0\) such that,
   \[ \|f(u)\| \leq k\|u\|, \text{ for every } u \in E. \]
4. The function \(f\) is continuous at every point of \(E\).

**Proof.** Assume (1). Then for every \(\epsilon > 0\), there is some \(\eta > 0\) such that, for every \(u \in E\), if \(\|u\| \leq \eta\), then \(\|f(u)\| \leq \epsilon\). Pick \(\epsilon = 1\), so that there is some \(\eta > 0\) such that, if \(\|u\| \leq \eta\), then \(\|f(u)\| \leq 1\). If \(\|u\| \leq 1\), then \(\|\eta u\| \leq \eta \|u\| \leq \eta\), and so, \(\|f(\eta u)\| \leq 1\), that is, \(\eta \|f(u)\| \leq 1\), which implies \(\|f(u)\| \leq \eta^{-1}\). Thus, (2) holds with \(k = \eta^{-1}\).

Assume that (2) holds. If \(u = 0\), then by linearity, \(f(0) = 0\), and thus \(\|f(0)\| \leq k\|0\|\) holds trivially for all \(k \geq 0\). If \(u \neq 0\), then \(\|u\| > 0\), and since
\[
\left\| \frac{u}{\|u\|} \right\| = 1,
\]
we have
\[
\left\| f\left( \frac{u}{\|u\|} \right) \right\| \leq k,
\]
which implies that
\[ \|f(u)\| \leq k\|u\|. \]
Thus, (3) holds.

If (3) holds, then for all \( u, v \in E \), we have
\[ \|f(v) - f(u)\| = \|f(v - u)\| \leq k\|v - u\|. \]
If \( k = 0 \), then \( f \) is the zero function, and continuity is obvious. Otherwise, if \( k > 0 \), for every \( \epsilon > 0 \), if \( \|v - u\| \leq \frac{\epsilon}{k} \), then \( \|f(v - u)\| \leq \epsilon \), which shows continuity at every \( u \in E \). Finally, it is obvious that (4) implies (1).

Among other things, Proposition 13.61 shows that a linear map is continuous iff the image of the unit (closed) ball is bounded. Since a continuous linear map satisfies the condition \( \|f(u)\| \leq k\|u\| \) (for some \( k \geq 0 \)), it is also uniformly continuous.

If \( E \) and \( F \) are normed vector spaces, the set of all continuous linear maps \( f : E \to F \) is denoted by \( \mathcal{L}(E; F) \).

Using Proposition 13.61, we can define a norm on \( \mathcal{L}(E; F) \) which makes it into a normed vector space. This definition has already been given in Chapter 14 (Definition 14.7) but for the reader’s convenience, we repeat it here.

**Definition 13.42.** Given two normed vector spaces \( E \) and \( F \), for every continuous linear map \( f : E \to F \), we define the *operator norm* \( \|f\| \) of \( f \) as
\[
\|f\| = \inf \{k \geq 0 \mid \|f(x)\| \leq k\|x\|, \text{ for all } x \in E\}
= \sup \{\|f(x)\| \mid \|x\| \leq 1\}
= \sup \{\|f(x)\| \mid \|x\| = 1\}.
\]

From Definition 13.42, for every continuous linear map \( f \in \mathcal{L}(E; F) \), we have
\[ \|f(x)\| \leq \|f\| \|x\|, \]
for every \( x \in E \). It is easy to verify that \( \mathcal{L}(E; F) \) is a normed vector space under the norm of Definition 13.42. Furthermore, if \( E, F, G \), are normed vector spaces, and \( f : E \to F \) and \( g : F \to G \) are continuous linear maps, we have
\[ \|g \circ f\| \leq \|g\| \|f\|. \]

We can now show that when \( E = \mathbb{R}^n \) or \( E = \mathbb{C}^n \), with any of the norms \( \|\|_1, \|\|_2, \) or \( \|\|_\infty \), then every linear map \( f : E \to F \) is continuous.

**Proposition 13.62.** If \( E = \mathbb{R}^n \) or \( E = \mathbb{C}^n \), with any of the norms \( \|\|_1, \|\|_2, \) or \( \|\|_\infty \), and \( F \) is any normed vector space, then every linear map \( f : E \to F \) is continuous.
Proof. Let \((e_1, \ldots, e_n)\) be the standard basis of \(\mathbb{R}^n\) (a similar proof applies to \(\mathbb{C}^n\)). In view of Proposition 14.2, it is enough to prove the proposition for the norm
\[
\|x\|_\infty = \max\{|x_i| \mid 1 \leq i \leq n\}.
\]
We have,
\[
\|f(v) - f(u)\| = \|f(v - u)\| = \left\| f \left( \sum_{1 \leq i \leq n} (v_i - u_i)e_i \right) \right\| = \left\| \sum_{1 \leq i \leq n} (v_i - u_i)f(e_i) \right\|,
\]
and so,
\[
\|f(v) - f(u)\| \leq \left( \sum_{1 \leq i \leq n} \|f(e_i)\| \right) \max_{1 \leq i \leq n} |v_i - u_i| = \left( \sum_{1 \leq i \leq n} \|f(e_i)\| \right) \|v - u\|_\infty.
\]
By the argument used in Proposition 13.61 to prove that (3) implies (4), \(f\) is continuous.

Actually, we proved in Theorem 14.3 that if \(E\) is a vector space of finite dimension, then any two norms are equivalent, so that they define the same topology. This fact together with Proposition 13.62 prove the following:

**Theorem 13.63.** If \(E\) is a vector space of finite dimension (over \(\mathbb{R}\) or \(\mathbb{C}\)), then all norms are equivalent (define the same topology). Furthermore, for any normed vector space \(F\), every linear map \(f : E \to F\) is continuous.

If \(E\) is a normed vector space of infinite dimension, a linear map \(f : E \to F\) may not be continuous. As an example, let \(E\) be the infinite vector space of all polynomials over \(\mathbb{R}\). Let
\[
\|P(X)\| = \max_{0 \leq x \leq 1} |P(x)|.
\]
We leave as an exercise to show that this is indeed a norm. Let \(F = \mathbb{R}\), and let \(f : E \to F\) be the map defined such that, \(f(P(X)) = P(3)\). It is clear that \(f\) is linear. Consider the sequence of polynomials
\[
P_n(X) = \left( \frac{X}{2} \right)^n.
\]
It is clear that \(\|P_n\| = \left( \frac{1}{2} \right)^n\), and thus, the sequence \(P_n\) has the null polynomial as a limit. However, we have
\[
f(P_n(X)) = P_n(3) = \left( \frac{3}{2} \right)^n,
\]
and the sequence \(f(P_n(X))\) diverges to \(+\infty\). Consequently, in view of Proposition 13.15 (1), \(f\) is not continuous.
We now consider the continuity of multilinear maps. We treat explicitly bilinear maps, the general case being a straightforward extension.

**Proposition 13.64.** Given normed vector spaces $E$, $F$ and $G$, for any bilinear map $f : E \times E \to G$, the following conditions are equivalent:

1. The function $f$ is continuous at $\langle 0, 0 \rangle$.

2. There is a constant $k \geq 0$ such that,
   \[ \|f(u, v)\| \leq k, \text{ for all } u, v \in E \text{ such that } \|u\|, \|v\| \leq 1. \]

3. There is a constant $k \geq 0$ such that,
   \[ \|f(u, v)\| \leq k\|u\|\|v\|, \text{ for all } u, v \in E. \]

4. The function $f$ is continuous at every point of $E \times F$.

**Proof.** It is similar to that of Proposition 13.61, with a small subtlety in proving that (3) implies (4), namely that two different $\eta$'s that are not independent are needed. \[ \square \]

In contrast to continuous linear maps, which must be uniformly continuous, nonzero continuous bilinear maps are not uniformly continuous. Let $f : E \times F \to G$ be a continuous bilinear map such that $f(a, b) \neq 0$ for some $a \in E$ and some $b \in F$. Consider the sequences $(u_n)$ and $(v_n)$ (with $n \geq 1$) given by

\[ u_n = (x_n, y_n) = (na, nb) \]
\[ v_n = (x'_n, y'_n) = \left( n + \frac{1}{n} \right) a, \left( n + \frac{1}{n} \right) b \].

Obviously
\[ \|v_n - u_n\| \leq \frac{1}{n}(\|a\| + \|b\|), \]
so $\lim_{n \to \infty} \|v_n - u_n\| = 0$. On the other hand
\[ f(x'_n, y'_n) - f(x_n, y_n) = \left( 2 + \frac{1}{n^2} \right) f(a, b), \]
and thus $\lim_{n \to \infty} \|f(x'_n, y'_n) - f(x_n, y_n)\| = 2 \|f(a, b)\| \neq 0$, which shows that $f$ is not uniformly continuous, because if this was the case, this limit would be zero.

If $E$, $F$, and $G$, are normed vector spaces, we denote the set of all continuous bilinear maps $f : E \times F \to G$ by $L_2(E,F;G)$. Using Proposition 13.64, we can define a norm on $L_2(E,F;G)$ which makes it into a normed vector space.
Definition 13.43. Given normed vector spaces $E$, $F$, and $G$, for every continuous bilinear map $f : E \times F \to G$, we define the norm $\|f\|$ of $f$ as

$$\|f\| = \inf \{k \geq 0 \mid \|f(x, y)\| \leq k \|x\| \|y\|, \text{ for all } x, y \in E\}$$

$$= \sup \{\|f(x, y)\| \mid \|x\|, \|y\| \leq 1\}.$$

From Definition 13.42, for every continuous bilinear map $f \in L_2(E, F; G)$, we have

$$\|f(x, y)\| \leq \|f\| \|x\| \|y\|,$$

for all $x, y \in E$. It is easy to verify that $L_2(E, F; G)$ is a normed vector space under the norm of Definition 13.43.

Given a bilinear map $f : E \times F \to G$, for every $u \in E$, we obtain a linear map denoted $fu : F \to G$, defined such that, $fu(v) = f(u, v)$. Furthermore, since

$$\|f(x, y)\| \leq \|f\| \|x\| \|y\|,$$

it is clear that $fu$ is continuous. We can then consider the map $\varphi : E \to L(F; G)$, defined such that, $\varphi(u) = fu$, for any $u \in E$, or equivalently, such that,

$$\varphi(u)(v) = f(u, v).$$

Actually, it is easy to show that $\varphi$ is linear and continuous, and that $\|\varphi\| = \|f\|$. Thus, $f \mapsto \varphi$ defines a map from $L_2(E, F; G)$ to $L(E; L(F; G))$. We can also go back from $L(E; L(F; G))$ to $L_2(E, F; G)$. We summarize all this in the following proposition.

Proposition 13.65. Let $E, F, G$ be three normed vector spaces. The map $f \mapsto \varphi$, from $L_2(E, F; G)$ to $L(E; L(F; G))$, defined such that, for every $f \in L_2(E, F; G)$,

$$\varphi(u)(v) = f(u, v),$$

is an isomorphism of vector spaces, and furthermore, $\|\varphi\| = \|f\|$.

As a corollary of Proposition 13.65, we get the following proposition which will be useful when we define second-order derivatives.

Proposition 13.66. Let $E, F$ be normed vector spaces. The map $app$ from $L(E; F) \times E$ to $F$, defined such that, for every $f \in L(E; F)$, for every $u \in E$,

$$app(f, u) = f(u),$$

is a continuous bilinear map.
Remark: If $E$ and $F$ are nontrivial, it can be shown that $\|\text{app}\| = 1$. It can also be shown that composition
\[
\circ: \mathcal{L}(E; F) \times \mathcal{L}(F; G) \rightarrow \mathcal{L}(E; G),
\]
is bilinear and continuous.

The above propositions and definition generalize to arbitrary $n$-multilinear maps, with $n \geq 2$. Proposition 13.64 extends in the obvious way to any $n$-multilinear map $f: E_1 \times \cdots \times E_n \rightarrow F$, but condition (3) becomes:

There is a constant $k \geq 0$ such that,
\[
\|f(u_1, \ldots, u_n)\| \leq k\|u_1\| \cdots \|u_n\|, \quad \text{for all } u_1 \in E_1, \ldots, u_n \in E_n.
\]

Definition 13.43 also extends easily to
\[
\|f\| = \inf \{k \geq 0 \mid \|f(x_1, \ldots, x_n)\| \leq k\|x_1\| \cdots \|x_n\|, \quad \text{for all } x_i \in E_i, 1 \leq i \leq n\}
\]
\[
= \sup \{\|f(x_1, \ldots, x_n)\| \mid \|x_n\|, \ldots, \|x_1\| \leq 1\}.
\]

Proposition 13.65 is also easily extended, and we get an isomorphism between continuous $n$-multilinear maps in $\mathcal{L}_n(E_1, \ldots, E_n; F)$, and continuous linear maps in
\[
\mathcal{L}(E_1; \mathcal{L}(E_2; \cdots; \mathcal{L}(E_n; F)))
\]

An obvious extension of Proposition 13.66 also holds.

Definition 13.44. A normed vector space $(E, \|\|)$ over $\mathbb{R}$ (or $\mathbb{C}$) which is a complete metric space for the distance $d(u, v) = \|v - u\|$, is called a Banach space.

It can be shown that every normed vector space of finite dimension is a Banach space (is complete). This is because $\mathbb{R}$ (and $\mathbb{C}$) are complete. The following theorem is a key result of the theory of Banach spaces worth proving.

**Theorem 13.67.** If $E$ and $F$ are normed vector spaces, and if $F$ is a Banach space, then $\mathcal{L}(E; F)$ is a Banach space (with the operator norm).

**Proof.** Let $(f)_{n \geq 1}$ be a Cauchy sequence of continuous linear maps $f_n: E \rightarrow F$. We proceed in several steps.

**Step 1.** Define the pointwise limit $f: E \rightarrow F$ of the sequence $(f_n)_{n \geq 1}$.

Since $(f)_{n \geq 1}$ is a Cauchy sequence, for every $\epsilon > 0$, there is some $N > 0$ such that $\|f_m - f_n\| < \epsilon$ for all $m, n \geq N$. Since $\|\|$ is the operator norm, we deduce that for any $u \in E$, we have
\[
\|f_m(u) - f_n(u)\| = \|(f_m - f_n)(u)\| \leq \|f_m - f_n\| \|u\| \leq \epsilon \|u\| \quad \text{for all } m, n \geq N,
\]
that is,
\[ \|f_m(u) - f_n(u)\| \leq \epsilon \|u\| \quad \text{for all } m, n \geq N. \] (\text{*}_1)

If \( u = 0 \), then \( f_m(0) = f_n(0) = 0 \) for all \( m, n \), so the sequence \( (f_n(0)) \) is a Cauchy sequence in \( F \) converging to 0. If \( u \neq 0 \), by replacing \( \epsilon \) by \( \epsilon / \|u\| \), we see that the sequence \( (f_n(u)) \) is a Cauchy sequence in \( F \). Since \( F \) is complete, the sequence \( (f_n(u)) \) has a limit which we denote by \( f(u) \). This defines our candidate limit function \( f \) by

\[ f(u) = \lim_{n \to \infty} f_n(u). \]

It remains to prove that

1. \( f \) is linear.
2. \( f \) is continuous.
3. \( f \) is the limit of \( (f_n) \) for the operator norm.

\textit{Step 2.} The function \( f \) is linear.

Recall that in a normed vector space, addition and multiplication by a fixed scalar are continuous (since \( \|u + v\| \leq \|u\| + \|v\| \) and \( \|\lambda u\| \leq |\lambda| \|u\| \)). Thus by definition of \( f \) and since the \( f_n \) are linear we have

\[
\begin{align*}
  f(u + v) &= \lim_{n \to \infty} f_n(u + v) & \text{by definition of } f \\
  &= \lim_{n \to \infty} (f_n(u) + f_n(v)) & \text{by linearity of } f_n \\
  &= \lim_{n \to \infty} f_n(u) + \lim_{n \to \infty} f_n(v) & \text{since } + \text{ is continuous} \\
  &= f(u) + f(v) & \text{by definition of } f.
\end{align*}
\]

Similarly,

\[
\begin{align*}
  f(\lambda u) &= \lim_{n \to \infty} f_n(\lambda u) & \text{by definition of } f \\
  &= \lim_{n \to \infty} \lambda f_n(u) & \text{by linearity of } f_n \\
  &= \lambda \lim_{n \to \infty} f_n(u) & \text{by continuity of scalar multiplication} \\
  &= \lambda f(u) & \text{by definition of } f.
\end{align*}
\]

Therefore, \( f \) is linear.

\textit{Step 3.} The function \( f \) is continuous.

Since \( (f_n)_{n \geq 1} \) is a Cauchy sequence, for every \( \epsilon > 0 \), there is some \( N > 0 \) such that \( \|f_m - f_n\| < \epsilon \) for all \( m, n \geq N \). Since \( f_m = f_n + f_m - f_n \), we get \( \|f_m\| \leq \|f_n\| + \|f_m - f_n\| \), which implies that

\[ \|f_m\| \leq \|f_n\| + \epsilon \quad \text{for all } m, n \geq N. \] (\text{*}_2)
CHAPTER 13. TOPOLOGY

Using \((\ast _2)\), we also have
\[
\| f_m(u) \| \leq \| f_m \| \| u \| \leq (\| f_n \| + \epsilon) \| u \| \quad \text{for all } m, n \geq N,
\]
that is,
\[
\| f_m(u) \| \leq (\| f_n \| + \epsilon) \| u \| \quad \text{for all } m, n \geq N. \tag{\ast _3}
\]
Hold \( n \geq N \) fixed and let \( m \) tend to \(+\infty\) in \((\ast _3)\). Since the norm is continuous, we get
\[
\| f(u) \| \leq (\| f_n \| + \epsilon) \| u \| ,
\]
which shows that \( f \) is continuous.

**Step 4.** The function \( f \) is the limit of \((f_n)\) for the operator norm.

Recall \((\ast _1)\):
\[
\| f_m(u) - f_n(u) \| \leq \epsilon \| u \| \quad \text{for all } m, n \geq N. \tag{\ast _1}
\]
Hold \( n \geq N \) fixed but this time let \( m \) tend to \(+\infty\) in \((\ast _1)\). By continuity of the norm we get
\[
\| f(u) - f_n(u) \| = \|(f - f_n)(u)\| \leq \epsilon \| u \| .
\]
By definition of the operator norm,
\[
\| f - f_n \| = \sup\{\|(f - f_n)(u)\| \mid \| u \| = 1\} \leq \epsilon \quad \text{for all } n \geq N,
\]
which proves that \( f_n \) converges to \( f \) for the operator norm.

As a special case of Theorem 13.67, if we let \( F = \mathbb{R} \) (or \( F = \mathbb{C} \) in the case of complex vector spaces) we see that \( E' = \mathcal{L}(E; \mathbb{R}) \) (or \( E' = \mathcal{L}(E; \mathbb{C}) \)) is complete (since \( \mathbb{R} \) and \( \mathbb{C} \) are complete). The space \( E' \) of continuous linear forms on \( E \) is called the dual of \( E \). It is a subspace of the algebraic dual \( E^* \) of \( E \) which consists of all linear forms on \( E \), not necessarily continuous.

It can also be shown that if \( E, F \) and \( G \) are normed vector spaces, and if \( G \) is a Banach space, then \( \mathcal{L}_2(E, F; G) \) is a Banach space. The proof is essentially identical.

### 13.12 Completion of a Normed Vector Space

An easy corollary of Theorem 13.58 and Theorem 13.57 is that every normed vector space can be embedded in a complete normed vector space, that is, a Banach space.

**Theorem 13.68.** If \((E, \| \|)\) is a normed vector space, then its completion \((\hat{E}, \hat{d})\) as a metric space (where \( E \) is given the metric \( d(x, y) = \| x - y \| \)) can be given a unique vector space structure extending the vector space structure on \( E \), and a norm \( \| \|_{\hat{E}} \), so that \((\hat{E}, \| \|_{\hat{E}})\) is a Banach space, and the metric \( \hat{d} \) is associated with the norm \( \| \|_{\hat{E}} \). Furthermore, the isometry \( \varphi : E \rightarrow \hat{E} \) is a linear isometry.
Proof. The addition operation $+ : E \times E \to E$ is uniformly continuous because

$$\| (u' + v') - (u'' + v'') \| \leq \| u' - u'' \| + \| v' - v'' \|.$$ 

It is not hard to show that $\hat{E} \times \hat{E}$ is a complete metric space and that $E \times E$ is dense in $\hat{E} \times \hat{E}$. Then, by Theorem 13.57, the uniformly continuous function $+$ has a unique continuous extension $+: \hat{E} \times \hat{E} \to \hat{E}$.

The map $\cdot : \mathbb{R} \times E \to E$ is not uniformly continuous, but for any fixed $\lambda \in \mathbb{R}$, the map $L_\lambda : E \to E$ given by $L_\lambda(u) = \lambda \cdot u$ is uniformly continuous, so by Theorem 13.57 the function $L_\lambda$ has a unique continuous extension $L_\lambda : \hat{E} \to \hat{E}$, which we use to define the scalar multiplication $\cdot : \mathbb{R} \times \hat{E} \to \hat{E}$. It is easily checked that with the above addition and scalar multiplication, $\hat{E}$ is a vector space.

Since the norm $\| \cdot \|$ on $E$ is uniformly continuous, it has a unique continuous extension $\| \| : \hat{E} \to \mathbb{R}_+$. The identities $\| u + v \| \leq \| u \| + \| v \|$ and $\| \lambda u \| \leq |\lambda| \| u \|$ extend to $\hat{E}$ by continuity. The equation

$$d(u, v) = \| u - v \|$$

also extends to $\hat{E}$ by continuity and yields

$$\hat{d}(\alpha, \beta) = \| \alpha - \beta \|_{\hat{E}},$$

which shows that $\| \|_{\hat{E}}$ is indeed a norm, and that the metric $\hat{d}$ is associated to it. Finally, it is easy to verify that the map $\varphi$ is linear. The uniqueness of the structure of normed vector space follows from the uniqueness of continuous extensions in Theorem 13.57.

Theorem 13.68 and Theorem 13.57 will be used to show that every Hermitian space can be embedded in a Hilbert space.

The following version of Theorem 13.57 for normed vector spaces will be needed in the theory of integration.

**Theorem 13.69.** Let $E$ and $F$ be two normed vector spaces, let $E_0$ be a dense subspace of $E$, and let $f_0 : E_0 \to F$ be a continuous function. If $f_0$ is uniformly continuous and if $F$ is complete, then there is a unique uniformly continuous function $f : E \to F$ extending $f_0$. Furthermore, if $f_0$ is a continuous linear map, then $f$ is also a linear continuous map, and $\| f \| = \| f_0 \|$.

**Proof.** We only need to prove the second statement. Given any two vectors $x, y \in E$, since $E_0$ is dense on $E$ we can pick sequences $(x_n)$ and $(y_n)$ of vectors $x_n, y_n \in E_0$ such that $x = \lim_{n \to \infty} x_n$ and $y = \lim_{n \to \infty} y_n$. Since addition and scalar multiplication are continuous, we get

$$x + y = \lim_{n \to \infty} (x_n + y_n),$$

$$\lambda x = \lim_{n \to \infty} (\lambda x_n).$$
for any $\lambda \in \mathbb{R}$ (or $\lambda \in \mathbb{C}$). Since $f(x)$ is defined by
\[ f(x) = \lim_{n \to \infty} f_0(x_n) \]
individually of the sequence $(x_n)$ converging to $x$, and similarly for $f(y)$ and $f(x+y)$, since $f_0$ is linear, we have
\[ f(x+y) = \lim_{n \to \infty} f_0(x_n + y_n) = \lim_{n \to \infty} \left( f_0(x_n) + f_0(y_n) \right) = \lim_{n \to \infty} f_0(x_n) + \lim_{n \to \infty} f_0(y_n) = f(x) + f(y). \]
Similarly,
\[ f(\lambda x) = \lim_{n \to \infty} f_0(\lambda x_n) = \lim_{n \to \infty} \lambda f_0(x_n) = \lambda \lim_{n \to \infty} f_0(x_n) = \lambda f(x). \]
Therefore, $f$ is linear. Since the norm is continuous, we have
\[ \|f(x)\| = \left\| \lim_{n \to \infty} f_0(x_n) \right\| = \lim_{n \to \infty} \|f_0(x_n)\|, \]
and since $f_0$ is continuous
\[ \|f_0(x_n)\| \leq \|f_0\| \|x_n\| \quad \text{for all } n \geq 1, \]
so we get
\[ \lim_{n \to \infty} \|f_0(x_n)\| \leq \lim_{n \to \infty} \|f_0\| \|x_n\| \quad \text{for all } n \geq 1, \]
that is,
\[ \|f(x)\| \leq \|f_0\| \|x\|. \]
Since
\[ \|f\| = \sup_{\|x\|=1, x \in E} \|f(x)\|, \]
we deduce that $\|f\| \leq \|f_0\|$. But since $E_0 \subseteq E$ and $f$ agrees with $f_0$ on $E_0$, we also have
\[ \|f_0\| = \sup_{\|x\|=1, x \in E_0} \|f_0(x)\| = \sup_{\|x\|=1, x \in E_0} \|f(x)\| \leq \sup_{\|x\|=1, x \in E} \|f(x)\| = \|f\|, \]
and thus $\|f\| = \|f_0\|$.

**13.13 Further Readings**

A thorough treatment of general topology can be found in Munkres [33, 32], Dixmier [13], Lang [25, 24], Schwartz [39, 38], and Bredon [9].
Chapter 14

Vector Norms and Matrix Norms

14.1 Normed Vector Spaces

In order to define how close two vectors or two matrices are, and in order to define the convergence of sequences of vectors or matrices, we can use the notion of a norm. Recall that $\mathbb{R}_+ = \{ x \in \mathbb{R} \mid x \geq 0 \}$. Also recall that if $z = a + ib \in \mathbb{C}$ is a complex number, with $a, b \in \mathbb{R}$, then $\overline{z} = a - ib$ and $|z| = \sqrt{z\overline{z}} = \sqrt{a^2 + b^2}$ ($|z|$ is the modulus of $z$).

Definition 14.1. Let $E$ be a vector space over a field $K$, where $K$ is either the field $\mathbb{R}$ of reals, or the field $\mathbb{C}$ of complex numbers. A norm on $E$ is a function $\| \cdot \| : E \to \mathbb{R}_+$, assigning a nonnegative real number $\|u\|$ to any vector $u \in E$, and satisfying the following conditions for all $x, y, z \in E$:

(N1) $\|x\| \geq 0$, and $\|x\| = 0$ iff $x = 0$. (positivity)

(N2) $\|\lambda x\| = |\lambda| \|x\|$. (homogeneity (or scaling))

(N3) $\|x + y\| \leq \|x\| + \|y\|$. (triangle inequality)

A vector space $E$ together with a norm $\| \cdot \|$ is called a normed vector space.

By (N2), setting $\lambda = -1$, we obtain

$$\|-x\| = \|(-1)x\| = |-1| \|x\| = \|x\|;$$

that is, $\|-x\| = \|x\|$. From (N3), we have

$$\|x\| = \|x - y + y\| \leq \|x - y\| + \|y\|,$$

which implies that

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

By exchanging $x$ and $y$ and using the fact that by (N2),

$$\|y - x\| = \|-(x - y)\| = \|x - y\|,$$
we also have
\[ \|y\| - \|x\| \leq \|x - y\|. \]

Therefore, \[ ||x| - |y|| \leq \|x - y\|, \quad \text{for all } x, y \in E. \] \hfill (\ast)

Observe that setting \( \lambda = 0 \) in (N2), we deduce that \( \|0\| = 0 \) without assuming (N1).

Then, by setting \( y = 0 \) in (\ast), we obtain
\[ ||x|| \leq \|x\|, \quad \text{for all } x \in E. \]

Therefore, the condition \( \|x\| \geq 0 \) in (N1) follows from (N2) and (N3), and (N1) can be replaced by the weaker condition

\( \text{(N1')} \) For all \( x \in E \), if \( \|x\| = 0 \) then \( x = 0 \),

A function \( ||\| : E \to \mathbb{R} \) satisfying axioms (N2) and (N3) is called a seminorm. From the above discussion, a seminorm also has the properties

\[ \|x\| \geq 0 \text{ for all } x \in E, \text{ and } \|0\| = 0. \]

However, there may be nonzero vectors \( x \in E \) such that \( \|x\| = 0 \). Let us give some examples of normed vector spaces.

**Example 14.1.**

1. Let \( E = \mathbb{R} \), and \( \|x\| = |x| \), the absolute value of \( x \).

2. Let \( E = \mathbb{C} \), and \( \|z\| = |z| \), the modulus of \( z \).

3. Let \( E = \mathbb{R}^n \) (or \( E = \mathbb{C}^n \)). There are three standard norms. For every \( (x_1, \ldots, x_n) \in E \), we have the norm \( \|x\|_1 \), defined such that,

\[ \|x\|_1 = |x_1| + \cdots + |x_n|, \]

we have the *Euclidean norm* \( \|x\|_2 \), defined such that,

\[ \|x\|_2 = \left( |x_1|^2 + \cdots + |x_n|^2 \right)^{\frac{1}{2}}, \]

and the *sup-norm* \( \|x\|_\infty \), defined such that,

\[ \|x\|_\infty = \max\{|x_i| \mid 1 \leq i \leq n\}. \]

More generally, we define the \( \ell_p \)-norm (for \( p \geq 1 \)) by

\[ \|x\|_p = \left( |x_1|^p + \cdots + |x_n|^p \right)^{1/p}. \]

There are other norms besides the \( \ell_p \)-norms. Here are some examples.
1. For $E = \mathbb{R}^2$,
   \[ \|(u_1, u_2)\| = |u_1| + 2|u_2|. \]

2. For $E = \mathbb{R}^2$,
   \[ \|(u_1, u_2)\| = \left( (u_1 + u_2)^2 + u_1^2 \right)^{1/2}. \]

3. For $E = \mathbb{C}^2$,
   \[ \|(u_1, u_2)\| = |u_1 + iu_2| + |u_1 - iu_2|. \]

The reader should check that they satisfy all the axioms of a norm.

Some work is required to show the triangle inequality for the $\ell_p$-norm.

**Proposition 14.1.** If $E$ is a finite-dimensional vector space over $\mathbb{R}$ or $\mathbb{C}$, for every real number $p \geq 1$, the $\ell_p$-norm is indeed a norm.

**Proof.** The cases $p = 1$ and $p = \infty$ are easy and left to the reader. If $p > 1$, then let $q > 1$ such that
\[ \frac{1}{p} + \frac{1}{q} = 1. \]
We will make use of the following fact: for all $\alpha, \beta \in \mathbb{R}$, if $\alpha, \beta \geq 0$, then
\[ \alpha \beta \leq \frac{\alpha^p}{p} + \frac{\beta^q}{q}. \] (*)
To prove the above inequality, we use the fact that the exponential function $t \mapsto e^t$ satisfies the following convexity inequality:
\[ e^{\theta x + (1-\theta)y} \leq \theta e^x + (1 - \theta)e^y, \]
for all $x, y \in \mathbb{R}$ and all $\theta$ with $0 \leq \theta \leq 1$.

Since the case $\alpha \beta = 0$ is trivial, let us assume that $\alpha > 0$ and $\beta > 0$. If we replace $\theta$ by $1/p$, $x$ by $p \log \alpha$ and $y$ by $q \log \beta$, then we get
\[ e^{\frac{1}{p}p \log \alpha + \frac{1}{q}q \log \beta} \leq \frac{1}{p}e^{p \log \alpha} + \frac{1}{q}e^{q \log \beta}, \]
which simplifies to
\[ \alpha \beta \leq \frac{\alpha^p}{p} + \frac{\beta^q}{q}, \]
as claimed.

We will now prove that for any two vectors $u, v \in E$, we have
\[ \sum_{i=1}^n |u_i v_i| \leq \|u\|_p \|v\|_q. \] (**)
Since the above is trivial if \( u = 0 \) or \( v = 0 \), let us assume that \( u \neq 0 \) and \( v \neq 0 \). Then, the inequality (*) with \( \alpha = |u_i|/\|u\|_p \) and \( \beta = |v_i|/\|v\|_q \) yields

\[
\frac{|u_i v_i|}{\|u\|_p \|v\|_q} \leq \frac{|u_i|^p}{p \|u\|_p^p} + \frac{|v_i|^q}{q \|v\|_q^q},
\]

for \( i = 1, \ldots, n \), and by summing up these inequalities, we get

\[
\sum_{i=1}^n |u_i v_i| \leq \|u\|_p \|v\|_q,
\]
as claimed. To finish the proof, we simply have to prove that property (N3) holds, since (N1) and (N2) are clear. Now, for \( i = 1, \ldots, n \), we can write

\[
(|u_i| + |v_i|)^p = |u_i|(|u_i| + |v_i|)^{p-1} + |v_i|(|u_i| + |v_i|)^{p-1},
\]
so that by summing up these equations we get

\[
\sum_{i=1}^n (|u_i| + |v_i|)^p = \sum_{i=1}^n |u_i|(|u_i| + |v_i|)^{p-1} + \sum_{i=1}^n |v_i|(|u_i| + |v_i|)^{p-1},
\]
and using the inequality (**), we get

\[
\sum_{i=1}^n (|u_i| + |v_i|)^p \leq (\|u\|_p + \|v\|_p) \left( \sum_{i=1}^n (|u_i| + |v_i|)^{(p-1)q} \right)^{1/q}.
\]

However, \( 1/p + 1/q = 1 \) implies \( pq = p + q \), that is, \( (p-1)q = p \), so we have

\[
\sum_{i=1}^n (|u_i| + |v_i|)^p \leq (\|u\|_p + \|v\|_p) \left( \sum_{i=1}^n (|u_i| + |v_i|)^p \right)^{1/q},
\]
which yields

\[
\left( \sum_{i=1}^n (|u_i| + |v_i|)^p \right)^{1/p} \leq \|u\|_p + \|v\|_p,
\]
Since \( |u_i + v_i| \leq |u_i| + |v_i| \), the above implies the triangle inequality \( \|u + v\|_p \leq \|u\|_p + \|v\|_p \), as claimed.

For \( p > 1 \) and \( 1/p + 1/q = 1 \), the inequality

\[
\sum_{i=1}^n |u_i v_i| \leq \left( \sum_{i=1}^n |u_i|^p \right)^{1/p} \left( \sum_{i=1}^n |v_i|^q \right)^{1/q}
\]
is known as Hölder’s inequality. For \( p = 2 \), it is the Cauchy–Schwarz inequality.
Actually, if we define the Hermitian inner product $\langle - , - \rangle$ on $\mathbb{C}^n$ by

$$\langle u, v \rangle = \sum_{i=1}^{n} u_i \overline{v_i},$$

where $u = (u_1, \ldots, u_n)$ and $v = (v_1, \ldots, v_n)$, then

$$|\langle u, v \rangle| \leq \sum_{i=1}^{n} |u_i| |v_i| = \sum_{i=1}^{n} |u_i v_i|,$$

so Hölder’s inequality implies the inequality

$$|\langle u, v \rangle| \leq \|u\|_p \|v\|_q$$

also called Hölder’s inequality, which, for $p = 2$ is the standard Cauchy–Schwarz inequality. The triangle inequality for the $\ell_p$-norm,

$$\left( \sum_{i=1}^{n} (|u_i + v_i|)^p \right)^{1/p} \leq \left( \sum_{i=1}^{n} |u_i|^p \right)^{1/p} + \left( \sum_{i=1}^{n} |v_i|^q \right)^{1/q},$$

is known as Minkowski’s inequality.

When we restrict the Hermitian inner product to real vectors, $u, v \in \mathbb{R}^n$, we get the Euclidean inner product

$$\langle u, v \rangle = \sum_{i=1}^{n} u_i v_i.$$

It is very useful to observe that if we represent (as usual) $u = (u_1, \ldots, u_n)$ and $v = (v_1, \ldots, v_n)$ (in $\mathbb{R}^n$) by column vectors, then their Euclidean inner product is given by

$$\langle u, v \rangle = u^\top v = v^\top u,$$

and when $u, v \in \mathbb{C}^n$, their Hermitian inner product is given by

$$\langle u, v \rangle = v^* u = \overline{u^* v}.$$

In particular, when $u = v$, in the complex case we get

$$\|u\|_2^2 = u^* u,$$

and in the real case, this becomes

$$\|u\|_2^2 = u^\top u.$$

As convenient as these notations are, we still recommend that you do not abuse them; the notation $\langle u, v \rangle$ is more intrinsic and still “works” when our vector space is infinite dimensional.

The following proposition is easy to show.
Proposition 14.2. The following inequalities hold for all $x \in \mathbb{R}^n$ (or $x \in \mathbb{C}^n$):

\[
\|x\|_{\infty} \leq \|x\|_1 \leq n \|x\|_{\infty},
\]

\[
\|x\|_{\infty} \leq \|x\|_2 \leq \sqrt{n} \|x\|_{\infty},
\]

\[
\|x\|_2 \leq \|x\|_1 \leq \sqrt{n} \|x\|_2.
\]

Proposition 14.2 is a special case of a very important result: in a finite-dimensional vector space, any two norms are equivalent.

Definition 14.2. Given any (real or complex) vector space $E$, two norms $\|\cdot\|_a$ and $\|\cdot\|_b$ are equivalent iff there exists some positive reals $C_1, C_2 > 0$, such that

\[
\|u\|_a \leq C_1 \|u\|_b \quad \text{and} \quad \|u\|_b \leq C_2 \|u\|_a, \quad \text{for all } u \in E.
\]

Given any norm $\|\cdot\|$ on a vector space of dimension $n$, for any basis $(e_1, \ldots, e_n)$ of $E$, observe that for any vector $x = x_1e_1 + \cdots + x_ne_n$, we have

\[
\|x\| = \|x_1e_1 + \cdots + x_ne_n\| \leq |x_1|\|e_1\| + \cdots + |x_n|\|e_n\| \leq C(|x_1| + \cdots + |x_n|) = C\|x\|_1,
\]

with $C = \max_{1 \leq i \leq n} \|e_i\|$ and

\[
\|x\|_1 = \|x_1e_1 + \cdots + x_ne_n\| = |x_1| + \cdots + |x_n|.
\]

The above implies that

\[
\|u\| - \|v\| \leq \|u - v\| \leq C \|u - v\|_1,
\]

which means that the map $u \mapsto \|u\|$ is continuous with respect to the norm $\|\cdot\|_1$.

Let $S_{1}^{n-1}$ be the unit sphere with respect to the norm $\|\cdot\|_1$, namely

\[
S_{1}^{n-1} = \{ x \in E \mid \|x\|_1 = 1 \}.
\]

Now, $S_{1}^{n-1}$ is a closed and bounded subset of a finite-dimensional vector space, so by Heine–Borel (or equivalently, by Bolzano–Weierstrass), $S_{1}^{n-1}$ is compact. On the other hand, it is a well known result of analysis that any continuous real-valued function on a nonempty compact set has a minimum and a maximum, and that they are achieved. Using these facts, we can prove the following important theorem:

Theorem 14.3. If $E$ is any real or complex vector space of finite dimension, then any two norms on $E$ are equivalent.

Proof. It is enough to prove that any norm $\|\cdot\|$ is equivalent to the 1-norm. We already proved that the function $x \mapsto \|x\|$ is continuous with respect to the norm $\|\cdot\|_1$ and we observed that the unit sphere $S_{1}^{n-1}$ is compact. Now, we just recalled that because the function $f: x \mapsto \|x\|$ is continuous and because $S_{1}^{n-1}$ is compact, the function $f$ has a minimum $m$ and a maximum
14.2. MATRIX NORMS

$M$, and because $\|x\|$ is never zero on $S_1^{n-1}$, we must have $m > 0$. Consequently, we just proved that if $\|x\|_1 = 1$, then

$$0 < m \leq \|x\| \leq M,$$

so for any $x \in E$ with $x \neq 0$, we get

$$m \leq \|x/\|x\|_1\| \leq M,$$

which implies

$$m \|x\|_1 \leq \|x\| \leq M \|x\|_1.$$

Since the above inequality holds trivially if $x = 0$, we just proved that $\|\|$ and $\|\|_1$ are equivalent, as claimed. \qed

Next, we will consider norms on matrices.

14.2 Matrix Norms

For simplicity of exposition, we will consider the vector spaces $M_n(\mathbb{R})$ and $M_n(\mathbb{C})$ of square $n \times n$ matrices. Most results also hold for the spaces $M_{m,n}(\mathbb{R})$ and $M_{m,n}(\mathbb{C})$ of rectangular $m \times n$ matrices. Since $n \times n$ matrices can be multiplied, the idea behind matrix norms is that they should behave “well” with respect to matrix multiplication.

**Definition 14.3.** A matrix norm $\|\|$ on the space of square $n \times n$ matrices in $M_n(K)$, with $K = \mathbb{R}$ or $K = \mathbb{C}$, is a norm on the vector space $M_n(K)$, with the additional property called submultiplicativity that

$$\|AB\| \leq \|A\| \|B\|,$$

for all $A, B \in M_n(K)$. A norm on matrices satisfying the above property is often called a submultiplicative matrix norm.

Since $I^2 = I$, from $\|I\| = \|I^2\| \leq \|I\|^2$, we get $\|I\| \geq 1$, for every matrix norm.

Before giving examples of matrix norms, we need to review some basic definitions about matrices. Given any matrix $A = (a_{ij}) \in M_{m,n}(\mathbb{C})$, the conjugate $\overline{A}$ of $A$ is the matrix such that

$$\overline{A}_{ij} = \overline{a_{ij}}, \quad 1 \leq i \leq m, \quad 1 \leq j \leq n.$$

The transpose of $A$ is the $n \times m$ matrix $A^\top$ such that

$$A^*_{ij} = a_{ji}, \quad 1 \leq i \leq m, \quad 1 \leq j \leq n.$$

The adjoint of $A$ is the $n \times m$ matrix $A^*$ such that

$$A^* = (A^\top) = (\overline{A})^\top.$$
When $A$ is a real matrix, $A^* = A^\top$. A matrix $A \in M_n(\mathbb{C})$ is Hermitian if
\[ A^* = A. \]
If $A$ is a real matrix ($A \in M_n(\mathbb{R})$), we say that $A$ is symmetric if
\[ A^\top = A. \]
A matrix $A \in M_n(\mathbb{C})$ is normal if
\[ AA^* = A^*A, \]
and if $A$ is a real matrix, it is normal if
\[ AA^\top = A^\top A. \]
A matrix $U \in M_n(\mathbb{C})$ is unitary if
\[ UU^* = U^*U = I. \]
A real matrix $Q \in M_n(\mathbb{R})$ is orthogonal if
\[ QQ^\top = Q^\top Q = I. \]

Given any matrix $A = (a_{ij}) \in M_n(\mathbb{C})$, the trace $\text{tr}(A)$ of $A$ is the sum of its diagonal elements
\[ \text{tr}(A) = a_{11} + \cdots + a_{nn}. \]
It is easy to show that the trace is a linear map, so that
\[ \text{tr}(\lambda A) = \lambda \text{tr}(A) \]
and
\[ \text{tr}(A + B) = \text{tr}(A) + \text{tr}(B). \]
Moreover, if $A$ is an $m \times n$ matrix and $B$ is an $n \times m$ matrix, it is not hard to show that
\[ \text{tr}(AB) = \text{tr}(BA). \]

We also review eigenvalues and eigenvectors. We content ourselves with definition involving matrices.

**Definition 14.4.** Given any square matrix $A \in M_n(\mathbb{C})$, a complex number $\lambda \in \mathbb{C}$ is an eigenvalue of $A$ if there is some nonzero vector $u \in \mathbb{C}^n$, such that
\[ Au = \lambda u. \]
If $\lambda$ is an eigenvalue of $A$, then the nonzero vectors $u \in \mathbb{C}^n$ such that $Au = \lambda u$ are called eigenvectors of $A$ associated with $\lambda$; together with the zero vector, these eigenvectors form a subspace of $\mathbb{C}^n$ denoted by $E_\lambda(A)$, and called the eigenspace associated with $\lambda$. 

Remark: Note that Definition 14.4 requires an eigenvector to be nonzero. A somewhat unfortunate consequence of this requirement is that the set of eigenvectors is not a subspace, since the zero vector is missing! On the positive side, whenever eigenvectors are involved, there is no need to say that they are nonzero. The fact that eigenvectors are nonzero is implicitly used in all the arguments involving them, so it seems safer (but perhaps not as elegant) to stipulate that eigenvectors should be nonzero.

If $A$ is a square real matrix $A \in M_n(\mathbb{R})$, then we restrict Definition 14.4 to real eigenvalues $\lambda \in \mathbb{R}$ and real eigenvectors. However, it should be noted that although every complex matrix always has at least some complex eigenvalue, a real matrix may not have any real eigenvalues. For example, the matrix

$$A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

has the complex eigenvalues $i$ and $-i$, but no real eigenvalues. Thus, typically, even for real matrices, we consider complex eigenvalues.

Observe that $\lambda \in \mathbb{C}$ is an eigenvalue of $A$

1. iff $Au = \lambda u$ for some nonzero vector $u \in \mathbb{C}^n$
2. iff $(\lambda I - A)u = 0$
3. iff the matrix $\lambda I - A$ defines a linear map which has a nonzero kernel, that is, iff $\lambda I - A$ not invertible.

However, it is a standard fact of linear algebra that $\lambda I - A$ is not invertible iff

$$\det(\lambda I - A) = 0.$$  

Now, $\det(\lambda I - A)$ is a polynomial of degree $n$ in the indeterminate $\lambda$, in fact, of the form

$$\lambda^n - \text{tr}(A)\lambda^{n-1} + \cdots + (-1)^n \det(A).$$

Thus, we see that the eigenvalues of $A$ are the zeros (also called roots) of the above polynomial. Since every complex polynomial of degree $n$ has exactly $n$ roots, counted with their multiplicity, we have the following definition:

**Definition 14.5.** Given any square $n \times n$ matrix $A \in M_n(\mathbb{C})$, the polynomial

$$\det(\lambda I - A) = \lambda^n - \text{tr}(A)\lambda^{n-1} + \cdots + (-1)^n \det(A)$$

is called the **characteristic polynomial** of $A$. The $n$ (not necessarily distinct) roots $\lambda_1, \ldots, \lambda_n$ of the characteristic polynomial are all the **eigenvalues** of $A$ and constitute the **spectrum** of $A$. We let

$$\rho(A) = \max_{1 \leq i \leq n} |\lambda_i|$$

be the largest modulus of the eigenvalues of $A$, called the **spectral radius** of $A$. 
Since the eigenvalue $\lambda_1, \ldots, \lambda_n$ of $A$ are the zeros of the polynomial
\[ \det(\lambda I - A) = \lambda^n - \text{tr}(A)\lambda^{n-1} + \cdots + (-1)^n \det(A), \]
we deduce that
\[ \text{tr}(A) = \lambda_1 + \cdots + \lambda_n, \]
\[ \det(A) = \lambda_1 \cdots \lambda_n. \]

**Proposition 14.4.** For any matrix norm $\| \|$ on $M_n(\mathbb{C})$ and for any square $n \times n$ matrix $A \in M_n(\mathbb{C})$, we have
\[ \rho(A) \leq \| A \|. \]

*Proof.* Let $\lambda$ be some eigenvalue of $A$ for which $|\lambda|$ is maximum, that is, such that $|\lambda| = \rho(A)$. If $u (\neq 0)$ is any eigenvector associated with $\lambda$ and if $U$ is the $n \times n$ matrix whose columns are all $u$, then $Au = \lambda u$ implies
\[ AU = \lambda U, \]
and since
\[ |\lambda| \| U \| = \| \lambda U \| = \| AU \| \leq \| A \| \| U \| \]
and $U \neq 0$, we have $\| U \| \neq 0$, and get
\[ \rho(A) = |\lambda| \leq \| A \|, \]
as claimed. \qed

Proposition 14.4 also holds for any real matrix norm $\| \|$ on $M_n(\mathbb{R})$ but the proof is more subtle and requires the notion of induced norm. We prove it after giving Definition 14.7.

Now, it turns out that if $A$ is a real $n \times n$ symmetric matrix, then the eigenvalues of $A$ are all real and there is some orthogonal matrix $Q$ such that
\[ A = Q \text{diag}(\lambda_1, \ldots, \lambda_n) Q^\top, \]
where $\text{diag}(\lambda_1, \ldots, \lambda_n)$ denotes the matrix whose only nonzero entries (if any) are its diagonal entries, which are the (real) eigenvalues of $A$. Similarly, if $A$ is a complex $n \times n$ Hermitian matrix, then the eigenvalues of $A$ are all real and there is some unitary matrix $U$ such that
\[ A = U \text{diag}(\lambda_1, \ldots, \lambda_n) U^*, \]
where $\text{diag}(\lambda_1, \ldots, \lambda_n)$ denotes the matrix whose only nonzero entries (if any) are its diagonal entries, which are the (real) eigenvalues of $A$.

We now return to matrix norms. We begin with the so-called *Frobenius norm*, which is just the norm $\| \|_2$ on $\mathbb{C}^{n^2}$, where the $n \times n$ matrix $A$ is viewed as the vector obtained by concatenating together the rows (or the columns) of $A$. The reader should check that for any $n \times n$ complex matrix $A = (a_{ij})$,
\[ \left( \sum_{i,j=1}^n |a_{ij}|^2 \right)^{1/2} = \sqrt{\text{tr}(A^\ast A)} = \sqrt{\text{tr}(AA^\ast)}. \]
14.2. MATRIX NORMS

Definition 14.6. The Frobenius norm \( \| \cdot \|_F \) is defined so that for every square \( n \times n \) matrix \( A \in \mathbb{M}_n(\mathbb{C}) \),
\[
\| A \|_F = \left( \sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{1/2} = \sqrt{\text{tr}(AA^*)} = \sqrt{\text{tr}(A^*A)}.
\]

The following proposition show that the Frobenius norm is a matrix norm satisfying other nice properties.

Proposition 14.5. The Frobenius norm \( \| \cdot \|_F \) on \( \mathbb{M}_n(\mathbb{C}) \) satisfies the following properties:

(1) It is a matrix norm; that is, \( \| AB \|_F \leq \| A \|_F \| B \|_F \), for all \( A, B \in \mathbb{M}_n(\mathbb{C}) \).

(2) It is unitarily invariant, which means that for all unitary matrices \( U, V \), we have
\[
\| A \|_F = \| UA \|_F = \| AV \|_F = \| UAV \|_F.
\]

(3) \( \sqrt{\rho(A^*A)} \leq \| A \|_F \leq \sqrt{n} \sqrt{\rho(A^*A)} \), for all \( A \in \mathbb{M}_n(\mathbb{C}) \).

Proof. (1) The only property that requires a proof is the fact \( \| AB \|_F \leq \| A \|_F \| B \|_F \). This follows from the Cauchy–Schwarz inequality:
\[
\| AB \|_F^2 = \sum_{i,j=1}^{n} \left| \sum_{k=1}^{n} a_{ik} b_{kj} \right|^2 \\
\leq \sum_{i,j=1}^{n} \left( \sum_{h=1}^{n} |a_{ih}|^2 \right) \left( \sum_{k=1}^{n} |b_{kj}|^2 \right) \\
= \left( \sum_{i,h=1}^{n} |a_{ih}|^2 \right) \left( \sum_{k,j=1}^{n} |b_{kj}|^2 \right) = \| A \|_F^2 \| B \|_F^2.
\]

(2) We have
\[
\| A \|_F^2 = \text{tr}(A^*A) = \text{tr}(VV^*A^*A) = \text{tr}(V^*A^*AV) = \| AV \|_F^2,
\]
and
\[
\| A \|_F^2 = \text{tr}(A^*A) = \text{tr}(A^*U^*UA) = \| UA \|_F^2.
\]
The identity
\[
\| A \|_F^2 = \| UAV \|_F^2
\]
follows from the previous two.

(3) It is well known that the trace of a matrix is equal to the sum of its eigenvalues. Furthermore, \( A^*A \) is symmetric positive semidefinite (which means that its eigenvalues are nonnegative), so \( \rho(A^*A) \) is the largest eigenvalue of \( A^*A \) and
\[
\rho(A^*A) \leq \text{tr}(A^*A) \leq n \rho(A^*A),
\]
which yields (3) by taking square roots. \( \square \)
Remark: The Frobenius norm is also known as the Hilbert-Schmidt norm or the Schur norm. So many famous names associated with such a simple thing!

We now give another method for obtaining matrix norms using subordinate norms. First, we need a proposition that shows that in a finite-dimensional space, the linear map induced by a matrix is bounded, and thus continuous.

**Proposition 14.6.** For every norm $\| \|$ on $\mathbb{C}^n$ (or $\mathbb{R}^n$), for every matrix $A \in M_n(\mathbb{C})$ (or $A \in M_n(\mathbb{R})$), there is a real constant $C_A \geq 0$, such that

$$\|Au\| \leq C_A \|u\|,$$

for every vector $u \in \mathbb{C}^n$ (or $u \in \mathbb{R}^n$ if $A$ is real).

**Proof.** For every basis $(e_1, \ldots, e_n)$ of $\mathbb{C}^n$ (or $\mathbb{R}^n$), for every vector $u = u_1e_1 + \cdots + u_ne_n$, we have

$$\|Au\| = \|u_1A(e_1) + \cdots + u_nA(e_n)\| \leq |u_1|\|A(e_1)\| + \cdots + |u_n|\|A(e_n)\| \leq C_1(|u_1| + \cdots + |u_n|) = C_1\|u\|_1,$$

where $C_1 = \max_{1 \leq i \leq n} \|A(e_i)\|$. By Theorem 14.3, the norms $\| \|$ and $\| \|_1$ are equivalent, so there is some constant $C_2 > 0$ so that $\|u\|_1 \leq C_2\|u\|$ for all $u$, which implies that

$$\|Au\| \leq C_A \|u\|,$$

where $C_A = C_1C_2$. \qed

Proposition 14.6 says that every linear map on a finite-dimensional space is bounded. This implies that every linear map on a finite-dimensional space is continuous. Actually, it is not hard to show that a linear map on a normed vector space $E$ is bounded iff it is continuous, regardless of the dimension of $E$.

Proposition 14.6 implies that for every matrix $A \in M_n(\mathbb{C})$ (or $A \in M_n(\mathbb{R})$),

$$\sup_{\substack{x \in \mathbb{C}^n \\ x \neq 0 \ \ : \ \ \|x\| \neq 0}} \frac{\|Ax\|}{\|x\|} \leq C_A.$$

Now, since $\|\lambda u\| = |\lambda|\|u\|$, for every nonzero vector $x$, we have

$$\frac{\|Ax\|}{\|x\|} = \frac{\|x\| \|A(x/\|x\|)\|}{\|x\|} = \|A(x/\|x\|)\|,$$

which implies that

$$\sup_{\substack{x \in \mathbb{C}^n \ \ x \neq 0 \ \ : \ \ \|x\| = 1}} \frac{\|Ax\|}{\|x\|} = \sup_{x \in \mathbb{C}^n} \|Ax\|.$$
Similarly
\[
\sup_{x \in \mathbb{C}^n, x \neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{x \in \mathbb{C}^n, \|x\| = 1} \|Ax\|.
\]

The above considerations justify the following definition.

**Definition 14.7.** If \( \| \cdot \| \) is any norm on \( \mathbb{C}^n \), we define the function \( \|\cdot\| \) on \( M_n(\mathbb{C}) \) by

\[
\|A\| = \sup_{x \in \mathbb{C}^n, x \neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{x \in \mathbb{C}^n, \|x\| = 1} \|Ax\|.
\]

The function \( A \mapsto \|A\| \) is called the **subordinate matrix norm or operator norm** induced by the norm \( \| \cdot \| \).

It is easy to check that the function \( A \mapsto \|A\| \) is indeed a norm, and by definition, it satisfies the property

\[
\|Ax\| \leq \|A\| \|x\|, \quad \text{for all } x \in \mathbb{C}^n.
\]

A norm \( \| \cdot \| \) on \( M_n(\mathbb{C}) \) satisfying the above property is said to be **subordinate** to the vector norm \( \| \cdot \| \) on \( \mathbb{C}^n \). As a consequence of the above inequality, we have

\[
\|ABx\| \leq \|A\| \|Bx\| \leq \|A\| \|B\| \|x\|,
\]

for all \( x \in \mathbb{C}^n \), which implies that

\[
\|AB\| \leq \|A\| \|B\| \quad \text{for all } A, B \in M_n(\mathbb{C}),
\]

showing that \( A \mapsto \|A\| \) is a matrix norm (it is submultiplicative).

Observe that the operator norm is also defined by

\[
\|A\| = \inf \{ \lambda \in \mathbb{R} \mid \|Ax\| \leq \lambda \|x\|, \text{ for all } x \in \mathbb{C}^n \}.
\]

Since the function \( x \mapsto \|Ax\| \) is continuous (because \( \|Ay\| - \|Ax\| \mid \leq \|Ay - Ax\| \leq C_A \|x - y\| \) and the unit sphere \( S^{n-1} = \{ x \in \mathbb{C}^n \mid \|x\| = 1 \} \) is compact, there is some \( x \in \mathbb{C}^n \) such that \( \|x\| = 1 \) and

\[
\|Ax\| = \|A\|.
\]

Equivalently, there is some \( x \in \mathbb{C}^n \) such that \( x \neq 0 \) and

\[
\|Ax\| = \|A\| \|x\|.
\]

The definition of an operator norm also implies that

\[
\|I\| = 1.
\]

The above shows that the Frobenius norm is not a subordinate matrix norm (why?). The notion of subordinate norm can be slightly generalized.
CHAPTER 14. VECTOR NORMS AND MATRIX NORMS

Definition 14.8. If $K = \mathbb{R}$ or $K = \mathbb{C}$, for any norm $\| \|$ on $M_{m,n}(K)$, and for any two norms $\| \|_a$ on $K^n$ and $\| \|_b$ on $K^m$, we say that the norm $\| \|$ is subordinate to the norms $\| \|_a$ and $\| \|_b$ if
\[
\|Ax\|_b \leq \|A\| \|x\|_a \quad \text{for all } A \in M_{m,n}(K) \text{ and all } x \in K^n.
\]

Remark: For any norm $\| \|$ on $\mathbb{C}^n$, we can define the function $\| \|_R$ on $M_n(\mathbb{R})$ by
\[
\|A\|_R = \sup_{x \in \mathbb{R}^n, x \neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\| = 1} \|Ax\|.
\]

The function $A \mapsto \|A\|_R$ is a matrix norm on $M_n(\mathbb{R})$, and
\[
\|A\|_R \leq \|A\|,
\]
for all real matrices $A \in M_n(\mathbb{R})$. However, it is possible to construct vector norms $\| \|$ on $\mathbb{C}^n$ and real matrices $A$ such that
\[
\|A\|_R < \|A\|.
\]

In order to avoid this kind of difficulties, we define subordinate matrix norms over $M_n(\mathbb{C})$. Luckily, it turns out that $\|A\|_R = \|A\|$ for the vector norms, $\| \|_1$, $\| \|_2$, and $\| \|_\infty$.

We now prove Proposition 14.4 for real matrix norms.

Proposition 14.7. For any matrix norm $\| \|$ on $M_n(\mathbb{R})$ and for any square $n \times n$ matrix $A \in M_n(\mathbb{R})$, we have
\[
\rho(A) \leq \|A\|.
\]

Proof. We follow the proof in Denis Serre’s book [42]. If $A$ is a real matrix, the problem is that the eigenvectors associated with the eigenvalue of maximum modulus may be complex. We use a trick based on the fact that for every matrix $A$ (real or complex),
\[
\rho(A^k) = (\rho(A))^k,
\]
which is left as an exercise (it can be shown that if $(\lambda_1, \ldots, \lambda_n)$ are the (not necessarily distinct) eigenvalues of $A$, then $(\lambda_1^k, \ldots, \lambda_n^k)$ are the eigenvalues of $A^k$, for $k \geq 1$).

Pick any complex matrix norm $\| \|_c$ on $\mathbb{C}^n$ (for example, the Frobenius norm, or any subordinate matrix norm induced by a norm on $\mathbb{C}^n$). The restriction of $\| \|_c$ to real matrices is a real norm that we also denote by $\| \|_c$. Now, by Theorem 14.3, since $M_n(\mathbb{R})$ has finite dimension $n^2$, there is some constant $C > 0$ so that
\[
\|B\|_c \leq C \|B\|, \quad \text{for all } B \in M_n(\mathbb{R}).
\]
Furthermore, for every $k \geq 1$ and for every real $n \times n$ matrix $A$, by Proposition 14.4, $\rho(A^k) \leq \|A^k\|_c$, and because $\| \|$ is a matrix norm, $\|A^k\| \leq \|A\|^k$, so we have
\[
(\rho(A))^k = \rho(A^k) \leq \|A^k\|_c \leq C \|A^k\| \leq C \|A\|^k,
\]
for all \( k \geq 1 \). It follows that

\[
\rho(A) \leq C^{1/k} \|A\|, \quad \text{for all} \quad k \geq 1.
\]

However because \( C > 0 \), we have \( \lim_{k \to \infty} C^{1/k} = 1 \) (we have \( \lim_{k \to \infty} \frac{1}{k} \log(C) = 0 \)). Therefore, we conclude that

\[
\rho(A) \leq \|A\|,
\]
as desired. \( \square \)

We now determine explicitly what are the subordinate matrix norms associated with the vector norms \( \|\cdot\|_1, \|\cdot\|_2, \) and \( \|\cdot\|_\infty \).

**Proposition 14.8.** For every square matrix \( A = (a_{ij}) \in M_n(\mathbb{C}) \), we have

\[
\|A\|_1 = \sup_{x \in \mathbb{C}^n, \|x\|_1 = 1} \|Ax\|_1 = \max_j \sum_{i=1}^n |a_{ij}|
\]

\[
\|A\|_\infty = \sup_{x \in \mathbb{C}^n, \|x\|_\infty = 1} \|Ax\|_\infty = \max_i \sum_{j=1}^n |a_{ij}|
\]

\[
\|A\|_2 = \sup_{x \in \mathbb{C}^n, \|x\|_2 = 1} \|Ax\|_2 = \sqrt{\rho(A^*A)} = \sqrt{\rho(AA^*)}.
\]

Furthermore, \( \|A^*\|_2 = \|A\|_2 \), the norm \( \|\cdot\|_2 \) is unitarily invariant, which means that

\[
\|A\|_2 = \|UAV\|_2
\]

for all unitary matrices \( U, V \), and if \( A \) is a normal matrix, then \( \|A\|_2 = \rho(A) \).

**Proof.** For every vector \( u \), we have

\[
\|Au\|_1 = \sum_i \left| \sum_j a_{ij} u_j \right| \leq \sum_j |u_j| \sum_i |a_{ij}| \leq \left( \max_j \sum_i |a_{ij}| \right) \|u\|_1,
\]

which implies that

\[
\|A\|_1 \leq \max_j \sum_{i=1}^n |a_{ij}|.
\]

It remains to show that equality can be achieved. For this let \( j_0 \) be some index such that

\[
\max_j \sum_i |a_{ij}| = \sum_i |a_{ij_0}|,
\]

and let \( u_i = 0 \) for all \( i \neq j_0 \) and \( u_{j_0} = 1 \).
In a similar way, we have
\[ \|Au\|_\infty = \max_i \left| \sum_j a_{ij}u_j \right| \leq \left( \max_i \sum_j |a_{ij}| \right) \|u\|_\infty, \]
which implies that
\[ \|A\|_\infty \leq \max_i \sum_{j=1}^n |a_{ij}|. \]
To achieve equality, let \( i_0 \) be some index such that
\[ \max_i \sum_j |a_{ij}| = \sum_j |a_{i_0j}|. \]
The reader should check that the vector given by
\[ u_j = \begin{cases} \frac{\pi_{i_0j}}{|a_{i_0j}|} & \text{if } a_{i_0j} \neq 0 \\ 1 & \text{if } a_{i_0j} = 0 \end{cases} \]
works.

We have
\[ \|A\|_2^2 = \sup_{x \in \mathbb{C}_n, x^*x = 1} \|Ax\|_2^2 = \sup_{x \in \mathbb{C}_n, x^*x = 1} x^*A^*Ax. \]
Since the matrix \( A^*A \) is symmetric, it has real eigenvalues and it can be diagonalized with respect to an orthogonal matrix. These facts can be used to prove that the function \( x \mapsto x^*A^*Ax \) has a maximum on the sphere \( x^*x = 1 \) equal to the largest eigenvalue of \( A^*A \), namely, \( \rho(A^*A) \). We postpone the proof until we discuss optimizing quadratic functions. Therefore,
\[ \|A\|_2 = \sqrt{\rho(A^*A)}. \]
Let use now prove that \( \rho(A^*A) = \rho(AA^*). \) First, assume that \( \rho(A^*A) > 0 \). In this case, there is some eigenvector \( u (\neq 0) \) such that
\[ A^*Au = \rho(A^*A)u, \]
and since \( \rho(A^*A) > 0 \), we must have \( Au \neq 0 \). Since \( Au \neq 0 \),
\[ AA^*(Au) = \rho(A^*A)Au \]
which means that \( \rho(A^*A) \) is an eigenvalue of \( AA^* \), and thus
\[ \rho(A^*A) \leq \rho(AA^*). \]
Because \( (A^*)^* = A \), by replacing \( A \) by \( A^* \), we get
\[ \rho(AA^*) \leq \rho(A^*A), \]
and so $\rho(A^*A) = \rho(AA^*)$.

If $\rho(A^*A) = 0$, then we must have $\rho(AA^*) = 0$, since otherwise by the previous reasoning we would have $\rho(A^*A) = \rho(AA^*) > 0$. Hence, in all case

$$\|A\|^2_2 = \rho(A^*A) = \rho(AA^*) = \|A^*\|^2_2.$$  

For any unitary matrices $U$ and $V$, it is an easy exercise to prove that $V^*A^*AV$ and $A^*A$ have the same eigenvalues, so

$$\|A\|^2_2 = \rho(A^*A) = \rho(V^*A^*AV) = \|AV\|^2_2,$$

and also

$$\|A\|^2_2 = \rho(A^*A) = \rho(A^*U^*UA) = \|UA\|^2_2.$$  

Finally, if $A$ is a normal matrix ($AA^* = A^*A$), it can be shown that there is some unitary matrix $U$ so that

$$A = UDU^*,$$

where $D = \text{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix consisting of the eigenvalues of $A$, and thus

$$A^*A = (UDU^*)^*UDU^* = U^*D^*U^*UDU^* = UD^*DU^*.$$  

However, $D^*D = \text{diag}(|\lambda_1|^2, \ldots, |\lambda_n|^2)$, which proves that

$$\rho(A^*A) = \rho(D^*D) = \max_i |\lambda_i|^2 = (\rho(A))^2,$$

so that $\|A\|_2 = \rho(A)$.  

The norm $\|A\|_2$ is often called the **spectral norm**. Observe that property (3) of proposition 14.5 says that

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{n} \|A\|_2,$$

which shows that the Frobenius norm is an upper bound on the spectral norm. The Frobenius norm is much easier to compute than the spectral norm.

The reader will check that the above proof still holds if the matrix $A$ is real, confirming the fact that $\|A\|_R = \|A\|$ for the vector norms $\|\|_1$, $\|\|_2$, and $\|\|_\infty$. It is also easy to verify that the proof goes through for rectangular matrices, with the same formulae. Similarly, the Frobenius norm is also a norm on rectangular matrices. For these norms, whenever $AB$ makes sense, we have

$$\|AB\| \leq \|A\| \|B\|.$$  

**Remark:** Let ($E, \|\|$) and ($F, \|\|$) be two normed vector spaces (for simplicity of notation, we use the same symbol $\|\|$ for the norms on $E$ and $F$; this should not cause any confusion).
Recall that a function $f : E \to F$ is continuous if for every $a \in E$, for every $\epsilon > 0$, there is some $\eta > 0$ such that for all $x \in E$,

$$\text{if} \quad \|x - a\| \leq \eta \quad \text{then} \quad \|f(x) - f(a)\| \leq \epsilon.$$ 

It is not hard to show that a linear map $f : E \to F$ is continuous iff there is some constant $C \geq 0$ such that

$$\|f(x)\| \leq C \|x\| \text{ for all } x \in E.$$ 

If so, we say that $f$ is bounded (or a linear bounded operator). We let $\mathcal{L}(E; F)$ denote the set of all continuous (equivalently, bounded) linear maps from $E$ to $F$. Then, we can define the operator norm (or subordinate norm) $\|\|$ on $\mathcal{L}(E; F)$ as follows: for every $f \in \mathcal{L}(E; F)$,

$$\|f\| = \sup_{x \in E, x \neq 0} \frac{\|f(x)\|}{\|x\|} = \sup_{x \in E, \|x\|=1} \|f(x)\|,$$

or equivalently by

$$\|f\| = \inf \{\lambda \in \mathbb{R} \mid \|f(x)\| \leq \lambda \|x\|, \text{ for all } x \in E\}.$$ 

It is not hard to show that the map $f \mapsto \|f\|$ is a norm on $\mathcal{L}(E; F)$ satisfying the property

$$\|f(x)\| \leq \|f\| \|x\|$$

for all $x \in E$, and that if $f \in \mathcal{L}(E; F)$ and $g \in \mathcal{L}(F; G)$, then

$$\|g \circ f\| \leq \|g\| \|f\|.$$ 

Operator norms play an important role in functional analysis, especially when the spaces $E$ and $F$ are complete.

The following proposition is needed when to deal with the condition number of a matrix.

**Proposition 14.9.** Let $\|\|$ be any matrix norm and let $B$ be a matrix such that $\|B\| < 1$.

1. If $\|\|$ is a subordinate matrix norm, then the matrix $I + B$ is invertible and

$$\|(I + B)^{-1}\| \leq \frac{1}{1 - \|B\|}.$$ 

2. If a matrix of the form $I + B$ is singular, then $\|B\| \geq 1$ for every matrix norm (not necessarily subordinate).

**Proof.** (1) Observe that $(I + B)u = 0$ implies $Bu = -u$, so

$$\|u\| = \|Bu\|.$$
Recall that
\[ \|Bu\| \leq \|B\| \|u\| \]
for every subordinate norm. Since \( \|B\| < 1 \), if \( u \neq 0 \), then
\[ \|Bu\| < \|u\|, \]
which contradicts \( \|u\| = \|Bu\| \). Therefore, we must have \( u = 0 \), which proves that \( I + B \) is injective, and thus bijective, i.e., invertible. Then, we have
\[ (I + B)^{-1} + B(I + B)^{-1} = (I + B)(I + B)^{-1} = I, \]
so we get
\[ (I + B)^{-1} = I - B(I + B)^{-1}, \]
which yields
\[ \|(I + B)^{-1}\| \leq 1 + \|B\| \|(I + B)^{-1}\|, \]
and finally,
\[ \|(I + B)^{-1}\| \leq \frac{1}{1 - \|B\|}. \]

(2) If \( I + B \) is singular, then \( -1 \) is an eigenvalue of \( B \), and by Proposition 14.4, we get \( \rho(B) \leq \|B\| \), which implies \( 1 \leq \rho(B) \leq \|B\|. \)

The following result is needed to deal with the convergence of sequences of powers of matrices.

**Proposition 14.10.** For every matrix \( A \in M_n(\mathbb{C}) \) and for every \( \epsilon > 0 \), there is some subordinate matrix norm \( \|\| \) such that
\[ \|A\| \leq \rho(A) + \epsilon. \]

*Proof.* By the Schur form, there exists some invertible matrix \( U \) and some upper triangular matrix \( T \) such that
\[ A = UTU^{-1}, \]
and say that
\[ T = \begin{pmatrix} \lambda_1 & t_{12} & t_{13} & \cdots & t_{1n} \\ 0 & \lambda_2 & t_{23} & \cdots & t_{2n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{n-1} & t_{n-1n} \\ 0 & 0 & \cdots & 0 & \lambda_n \end{pmatrix}, \]
where \( \lambda_1, \ldots, \lambda_n \) are the eigenvalues of \( A \). For every \( \delta \neq 0 \), define the diagonal matrix
\[ D_\delta = \text{diag}(1, \delta, \delta^2, \ldots, \delta^{n-1}), \]
and consider the matrix

\[(UD_δ)^{-1}A(UD_δ) = D_δ^{-1}TD_δ = \begin{pmatrix}
\lambda_1 & \delta t_{12} & \delta^2 t_{13} & \cdots & \delta^{n-1} t_{1n} \\
0 & \lambda_2 & \delta t_{23} & \cdots & \delta^{n-2} t_{2n} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & \lambda_{n-1} & \delta t_{n-1n} \\
0 & 0 & \cdots & 0 & \lambda_n
\end{pmatrix}.
\]

Now, define the function \(\|\|: M_n(\mathbb{C}) \rightarrow \mathbb{R}\) by

\[\|B\| = \|(UD_δ)^{-1}B(UD_δ)\|_\infty,
\]

for every \(B \in M_n(\mathbb{C})\). Then it is easy to verify that the above function is the matrix norm subordinate to the vector norm

\[v \mapsto \|(UD_δ)^{-1}v\|_\infty.
\]

Furthermore, for every \(\epsilon > 0\), we can pick \(δ\) so that

\[\sum_{j=i+1}^{n} |δ^{j-i} t_{ij}| \leq \epsilon, \quad 1 \leq i \leq n - 1,
\]

and by definition of the norm \(\|\|_\infty\), we get

\[\|A\| \leq \rho(A) + \epsilon,
\]

which shows that the norm that we have constructed satisfies the required properties.

\[\square\]

Note that equality is generally not possible; consider the matrix

\[A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},
\]

for which \(\rho(A) = 0 < \|A\|\), since \(A \neq 0\).
Chapter 15
Basics of Groups and Group Actions

This chapter gathers basics of the theory of groups and group actions.

15.1 Groups, Subgroups, Cosets

Definition 15.1. A group is a set $G$ equipped with a binary operation $\cdot : G \times G \rightarrow G$ that associates an element $a \cdot b \in G$ to every pair of elements $a, b \in G$, and having the following properties: $\cdot$ is associative, has an identity element $e \in G$, and every element in $G$ is invertible (w.r.t. $\cdot$). More explicitly, this means that the following equations hold for all $a, b, c \in G$:

(G1) $a \cdot (b \cdot c) = (a \cdot b) \cdot c$. (associativity);
(G2) $a \cdot e = e \cdot a = a$. (identity);
(G3) For every $a \in G$, there is some $a^{-1} \in G$ such that $a \cdot a^{-1} = a^{-1} \cdot a = e$. (inverse).

A group $G$ is abelian (or commutative) if

$$a \cdot b = b \cdot a \quad \text{for all } a, b \in G.$$  

A set $M$ together with an operation $\cdot : M \times M \rightarrow M$ and an element $e$ satisfying only Conditions (G1) and (G2) is called a monoid. For example, the set $\mathbb{N} = \{0, 1, \ldots, n, \ldots\}$ of natural numbers is a (commutative) monoid under addition. However, it is not a group.

Some examples of groups are given below.

Example 15.1.

1. The set $\mathbb{Z} = \{\ldots, -n, \ldots, -1, 0, 1, \ldots, n, \ldots\}$ of integers is an abelian group under addition, with identity element 0. However, $\mathbb{Z}^* = \mathbb{Z} - \{0\}$ is not a group under multiplication.

263
2. The set \( \mathbb{Q} \) of rational numbers (fractions \( p/q \) with \( p, q \in \mathbb{Z} \) and \( q \neq 0 \)) is an abelian group under addition, with identity element 0. The set \( \mathbb{Q}^* = \mathbb{Q} - \{0\} \) is also an abelian group under multiplication, with identity element 1.

3. Given any nonempty set \( S \), the set of bijections \( f : S \to S \), also called permutations of \( S \), is a group under function composition (i.e., the multiplication of \( f \) and \( g \) is the composition \( g \circ f \)), with identity element the identity function \( \text{id}_S \). This group is not abelian as soon as \( S \) has more than two elements. The permutation group of the set \( S = \{1, \ldots, n\} \) is often denoted \( \mathfrak{S}_n \) and called the symmetric group on \( n \) elements.

4. For any positive integer \( p \in \mathbb{N} \), define a relation on \( \mathbb{Z} \), denoted \( m \equiv n \pmod{p} \), as follows:

\[
m \equiv n \pmod{p} \iff m - n = kp \text{ for some } k \in \mathbb{Z}.
\]

The reader will easily check that this is an equivalence relation, and, moreover, it is compatible with respect to addition and multiplication, which means that if \( m_1 \equiv n_1 \pmod{p} \) and \( m_2 \equiv n_2 \pmod{p} \), then \( m_1 + m_2 \equiv n_1 + n_2 \pmod{p} \) and \( m_1 m_2 \equiv n_1 n_2 \pmod{p} \). Consequently, we can define an addition operation and a multiplication operation of the set of equivalence classes \( \pmod{p} \):

\[
[m] + [n] = [m + n]
\]

and

\[
[m] \cdot [n] = [mn].
\]

The reader will easily check that addition of residue classes \( \pmod{p} \) induces an abelian group structure with \([0]\) as zero. This group is denoted \( \mathbb{Z}/p\mathbb{Z} \).

5. The set of \( n \times n \) invertible matrices with real (or complex) coefficients is a group under matrix multiplication, with identity element the identity matrix \( I_n \). This group is called the general linear group and is usually denoted by \( GL(n, \mathbb{R}) \) (or \( GL(n, \mathbb{C}) \)).

6. The set of \( n \times n \) invertible matrices \( A \) with real (or complex) coefficients such that \( \det(A) = 1 \) is a group under matrix multiplication, with identity element the identity matrix \( I_n \). This group is called the special linear group and is usually denoted by \( SL(n, \mathbb{R}) \) (or \( SL(n, \mathbb{C}) \)).

7. The set of \( n \times n \) matrices \( Q \) with real coefficients such that

\[
QQ^\top = Q^\top Q = I_n
\]

is a group under matrix multiplication, with identity element the identity matrix \( I_n \); we have \( Q^{-1} = Q^\top \). This group is called the orthogonal group and is usually denoted by \( O(n) \).
8. The set of \(n \times n\) invertible matrices \(Q\) with real coefficients such that
\[
QQ^\top = Q^\top Q = I_n \quad \text{and} \quad \det(Q) = 1
\]
is a group under matrix multiplication, with identity element the identity matrix \(I_n\);
as in (6), we have \(Q^{-1} = Q^\top\). This group is called the \textit{special orthogonal group} or \textit{rotation group} and is usually denoted by \(\text{SO}(n)\).

The groups in (5)–(8) are nonabelian for \(n \geq 2\), except for \(\text{SO}(2)\) which is abelian (but \(\text{O}(2)\) is not abelian).

It is customary to denote the operation of an abelian group \(G\) by +, in which case the inverse \(a^{-1}\) of an element \(a \in G\) is denoted by \(-a\).

The identity element of a group is \textit{unique}. In fact, we can prove a more general fact:

\textbf{Proposition 15.1.} \textit{If a binary operation }\cdot:\ M \times M \to M\text{ is associative and if }e' \in M\text{ is a left identity and }e'' \in M\text{ is a right identity, which means that }
\begin{align*}
e' \cdot a &= a \quad \text{for all } a \in M \quad \text{(G2l)} \\
a \cdot e'' &= a \quad \text{for all } a \in M, \quad \text{(G2r)}
\end{align*}
\textit{then }e' = e''\text{.}

\textit{Proof.} If we let \(a = e''\) in equation (G2l), we get
\[e' \cdot e'' = e'',\]
and if we let \(a = e'\) in equation (G2r), we get
\[e' \cdot e'' = e',\]
and thus
\[e' = e' \cdot e'' = e'',\]
as claimed. \(\square\)

Proposition 15.1 implies that the identity element of a monoid is unique, and since every group is a monoid, the identity element of a group is unique. Furthermore, every element in a group has a \textit{unique inverse}. This is a consequence of a slightly more general fact:

\textbf{Proposition 15.2.} \textit{In a monoid }\(M\) \textit{with identity element }\(e\), \textit{if some element }\(a \in M\) \textit{has some left inverse }\(a' \in M\) \textit{and some right inverse }\(a'' \in M\), \textit{which means that }
\begin{align*}
a' \cdot a &= e \quad \text{(G3l)} \\
a \cdot a'' &= e \quad \text{(G3r)}
\end{align*}
\textit{then }a' = a''.
Proof. Using (G3l) and the fact that $e$ is an identity element, we have

$$(a' \cdot a) \cdot a'' = e \cdot a'' = a''.$$ 

Similarly, Using (G3r) and the fact that $e$ is an identity element, we have

$$a' \cdot (a \cdot a'') = a' \cdot e = a'.$$

However, since $M$ is monoid, the operation $\cdot$ is associative, so

$$a' = a' \cdot (a \cdot a'') = (a' \cdot a) \cdot a'' = a'',$$

as claimed. \qed

Remark: Axioms (G2) and (G3) can be weakened a bit by requiring only (G2r) (the existence of a right identity) and (G3r) (the existence of a right inverse for every element) (or (G2l) and (G3l)). It is a good exercise to prove that the group axioms (G2) and (G3) follow from (G2r) and (G3r).

Definition 15.2. If a group $G$ has a finite number $n$ of elements, we say that $G$ is a group of order $n$. If $G$ is infinite, we say that $G$ has infinite order. The order of a group is usually denoted by $|G|$ (if $G$ is finite).

Given a group $G$, for any two subsets $R, S \subseteq G$, we let

$$RS = \{r \cdot s \mid r \in R, s \in S\}.$$ 

In particular, for any $g \in G$, if $R = \{g\}$, we write

$$gS = \{g \cdot s \mid s \in S\},$$

and similarly, if $S = \{g\}$, we write

$$Rg = \{r \cdot g \mid r \in R\}.$$ 

From now on, we will drop the multiplication sign and write $g_1g_2$ for $g_1 \cdot g_2$.

Definition 15.3. Let $G$ be a group. For any $g \in G$, define $L_g$, the left translation by $g$, by $L_g(a) = ga$, for all $a \in G$, and $R_g$, the right translation by $g$, by $R_g(a) = ag$, for all $a \in G$.

The following simple fact is often used.

Proposition 15.3. Given a group $G$, the translations $L_g$ and $R_g$ are bijections.

Proof. We show this for $L_g$, the proof for $R_g$ being similar.

If $L_g(a) = L_g(b)$, then $ga = gb$, and multiplying on the left by $g^{-1}$, we get $a = b$, so $L_g$ injective. For any $b \in G$, we have $L_g(g^{-1}b) = g g^{-1}b = b$, so $L_g$ is surjective. Therefore, $L_g$ is bijective. \qed
**Definition 15.4.** Given a group $G$, a subset $H$ of $G$ is a *subgroup of $G$* iff

1. The identity element $e$ of $G$ also belongs to $H$ ($e \in H$);
2. For all $h_1, h_2 \in H$, we have $h_1 h_2 \in H$;
3. For all $h \in H$, we have $h^{-1} \in H$.

The proof of the following proposition is left as an exercise.

**Proposition 15.4.** Given a group $G$, a subset $H \subseteq G$ is a subgroup of $G$ iff $H$ is nonempty and whenever $h_1, h_2 \in H$, then $h_1 h_2^{-1} \in H$.

If the group $G$ is finite, then the following criterion can be used.

**Proposition 15.5.** Given a finite group $G$, a subset $H \subseteq G$ is a subgroup of $G$ iff

1. $e \in H$;
2. $H$ is closed under multiplication.

**Proof.** We just have to prove that Condition (3) of Definition 15.4 holds. For any $a \in H$, since the left translation $L_a$ is bijective, its restriction to $H$ is injective, and since $H$ is finite, it is also bijective. Since $e \in H$, there is a unique $b \in H$ such that $L_a(b) = ab = e$. However, if $a^{-1}$ is the inverse of $a$ in $G$, we also have $L_a(a^{-1}) = aa^{-1} = e$, and by injectivity of $L_a$, we have $a^{-1} = b \in H$.

**Example 15.2.**

1. For any integer $n \in \mathbb{Z}$, the set
   
   $$n\mathbb{Z} = \{nk \mid k \in \mathbb{Z}\}$$

   is a subgroup of the group $\mathbb{Z}$.

2. The set of matrices
   
   $$\text{GL}^+(n, \mathbb{R}) = \{A \in \text{GL}(n, \mathbb{R}) \mid \det(A) > 0\}$$

   is a subgroup of the group $\text{GL}(n, \mathbb{R})$.

3. The group $\text{SL}(n, \mathbb{R})$ is a subgroup of the group $\text{GL}(n, \mathbb{R})$.

4. The group $\text{O}(n)$ is a subgroup of the group $\text{GL}(n, \mathbb{R})$.

5. The group $\text{SO}(n)$ is a subgroup of the group $\text{O}(n)$, and a subgroup of the group $\text{SL}(n, \mathbb{R})$. 
6. It is not hard to show that every $2 \times 2$ rotation matrix $R \in \text{SO}(2)$ can be written as
\[
R = \begin{pmatrix}
\cos \theta & -\sin \theta \\
\sin \theta & \cos \theta
\end{pmatrix}, \quad \text{with} \quad 0 \leq \theta < 2\pi.
\]
Then $\text{SO}(2)$ can be considered as a subgroup of $\text{SO}(3)$ by viewing the matrix
\[
R = \begin{pmatrix}
\cos \theta & -\sin \theta \\
\sin \theta & \cos \theta
\end{pmatrix}
\]
as the matrix
\[
Q = \begin{pmatrix}
\cos \theta & -\sin \theta & 0 \\
\sin \theta & \cos \theta & 0 \\
0 & 0 & 1
\end{pmatrix}.
\]

7. The set of $2 \times 2$ upper-triangular matrices of the form
\[
\begin{pmatrix}
a & b \\
0 & c
\end{pmatrix}, \quad a, b, c \in \mathbb{R}, \quad a, c \neq 0
\]
is a subgroup of the group $\text{GL}(2, \mathbb{R})$.

8. The set $V$ consisting of the four matrices
\[
\begin{pmatrix}
\pm 1 & 0 \\
0 & \pm 1
\end{pmatrix}
\]
is a subgroup of the group $\text{GL}(2, \mathbb{R})$ called the *Klein four-group*.

**Definition 15.5.** If $H$ is a subgroup of $G$ and $g \in G$ is any element, the sets of the form $gH$ are called *left cosets of $H$ in $G$* and the sets of the form $Hg$ are called *right cosets of $H$ in $G*$. The left cosets (resp. right cosets) of $H$ induce an equivalence relation $\sim$ defined as follows: For all $g_1, g_2 \in G$,
\[
g_1 \sim g_2 \quad \text{iff} \quad g_1H = g_2H
\]
(resp. $g_1 \sim g_2$ iff $Hg_1 = Hg_2$). Obviously, $\sim$ is an equivalence relation.

Now, we claim the following fact:

**Proposition 15.6.** Given a group $G$ and any subgroup $H$ of $G$, we have $g_1H = g_2H$ iff $g_2^{-1}g_1H = H$ iff $g_2^{-1}g_1 \in H$, for all $g_1, g_2 \in G$.

**Proof.** If we apply the bijection $L_{g_2^{-1}}$ to both $g_1H$ and $g_2H$ we get $L_{g_2^{-1}}(g_1H) = g_2^{-1}g_1H$ and $L_{g_2^{-1}}(g_2H) = H$, so $g_1H = g_2H$ iff $g_2^{-1}g_1H = H$. If $g_2^{-1}g_1H = H$, since $1 \in H$, we get $g_2^{-1}g_1 \in H$. Conversely, if $g_2^{-1}g_1 \in H$, since $H$ is a group, the left translation $L_{g_2^{-1}g_1}$ is a bijection of $H$, so $g_2^{-1}g_1H = H$. Thus, $g_2^{-1}g_1H = H$ iff $g_2^{-1}g_1 \in H$. \qed
It follows that the equivalence class of an element \( g \in G \) is the coset \( gH \) (resp. \( Hg \)). Since \( L_g \) is a bijection between \( H \) and \( gH \), the cosets \( gH \) all have the same cardinality. The map \( L_{g^{-1}} \circ R_g \) is a bijection between the left coset \( gH \) and the right coset \( Hg \), so they also have the same cardinality. Since the distinct cosets \( gH \) form a partition of \( G \), we obtain the following fact:

**Proposition 15.7.** (Lagrange) For any finite group \( G \) and any subgroup \( H \) of \( G \), the order \( h \) of \( H \) divides the order \( n \) of \( G \).

**Definition 15.6.** Given a finite group \( G \) and a subgroup \( H \) of \( G \), if \( n = |G| \) and \( h = |H| \), then the ratio \( n/h \) is denoted by \((G:H)\) and is called the index of \( H \) in \( G \).

The index \((G:H)\) is the number of left (and right) cosets of \( H \) in \( G \). Proposition 15.7 can be stated as

\[ |G| = (G:H)|H|. \]

The set of left cosets of \( H \) in \( G \) (which, in general, is not a group) is denoted \( G/H \). The “points” of \( G/H \) are obtained by “collapsing” all the elements in a coset into a single element.

**Example 15.3.**

1. Let \( n \) be any positive integer, and consider the subgroup \( n\mathbb{Z} \) of \( \mathbb{Z} \) (under addition).

   The coset of 0 is the set \{0\}, and the coset of any nonzero integer \( m \in \mathbb{Z} \) is

   \[ m + n\mathbb{Z} = \{m + nk \mid k \in \mathbb{Z}\}. \]

   By dividing \( m \) by \( n \), we have \( m = nq + r \) for some unique \( r \) such that \( 0 \leq r \leq n - 1 \). But then we see that \( r \) is the smallest positive element of the coset \( m + n\mathbb{Z} \). This implies that there is a bijection between the cosets of the subgroup \( n\mathbb{Z} \) of \( \mathbb{Z} \) and the set of residues \{0, 1, \ldots, n - 1\} modulo \( n \), or equivalently a bijection with \( \mathbb{Z}/n\mathbb{Z} \).

2. The cosets of \( \text{SL}(n, \mathbb{R}) \) in \( \text{GL}(n, \mathbb{R}) \) are the sets of matrices

   \[ A\text{SL}(n, \mathbb{R}) = \{AB \mid B \in \text{SL}(n, \mathbb{R})\}, A \in \text{GL}(n, \mathbb{R}). \]

   Since \( A \) is invertible, \( \det(A) \neq 0 \), and we can write \( A = (\det(A))^{1/n}(\det(A))^{-1/n}A \) if \( \det(A) > 0 \) and \( A = (-\det(A))^{1/n}((-\det(A))^{-1/n}A) \) if \( \det(A) < 0 \). But we have \( (\det(A))^{-1/n}A \in \text{SL}(n, \mathbb{R}) \) if \( \det(A) > 0 \) and \( -(-\det(A))^{-1/n}A \in \text{SL}(n, \mathbb{R}) \) if \( \det(A) < 0 \), so the coset \( A\text{SL}(n, \mathbb{R}) \) contains the matrix

   \[ (\det(A))^{1/n}I_n \quad \text{if} \quad \det(A) > 0, \quad -(-\det(A))^{1/n}I_n \quad \text{if} \quad \det(A) < 0. \]

   It follows that there is a bijection between the cosets of \( \text{SL}(n, \mathbb{R}) \) in \( \text{GL}(n, \mathbb{R}) \) and \( \mathbb{R} \).
3. The cosets of $\text{SO}(n)$ in $\text{GL}^+(n, \mathbb{R})$ are the sets of matrices

$$A \text{SO}(n) = \{AQ \mid Q \in \text{SO}(n)\}, \ A \in \text{GL}^+(n, \mathbb{R}).$$

It can be shown (using the polar form for matrices) that there is a bijection between the cosets of $\text{SO}(n)$ in $\text{GL}^+(n, \mathbb{R})$ and the set of $n \times n$ symmetric, positive, definite matrices; these are the symmetric matrices whose eigenvalues are strictly positive.

4. The cosets of $\text{SO}(2)$ in $\text{SO}(3)$ are the sets of matrices

$$Q \text{SO}(2) = \{QR \mid R \in \text{SO}(2)\}, \ Q \in \text{SO}(3).$$

The group $\text{SO}(3)$ moves the points on the sphere $S^2$ in $\mathbb{R}^3$, namely for any $x \in S^2$,

$$x \mapsto Qx \quad \text{for any rotation } Q \in \text{SO}(3).$$

Here,

$$S^2 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}.$$

Let $N = (0, 0, 1)$ be the north pole on the sphere $S^2$. Then it is not hard to show that $\text{SO}(2)$ is precisely the subgroup of $\text{SO}(3)$ that leaves $N$ fixed. As a consequence, all rotations $QR$ in the coset $Q \text{SO}(2)$ map $N$ to the same point $QN \in S^2$, and it can be shown that there is a bijection between the cosets of $\text{SO}(2)$ in $\text{SO}(3)$ and the points on $S^2$. The surjectivity of this map has to do with the fact that the action of $\text{SO}(3)$ on $S^2$ is transitive, which means that for any point $x \in S^2$, there is some rotation $Q \in \text{SO}(3)$ such that $QN = x$.

It is tempting to define a multiplication operation on left cosets (or right cosets) by setting

$$(g_1H)(g_2H) = (g_1g_2)H,$$

but this operation is not well defined in general, unless the subgroup $H$ possesses a special property. In Example 15.3, it is possible to define multiplication of cosets in (1), but it is not possible in (2) and (3).

The property of the subgroup $H$ that allows defining a multiplication operation on left cosets is typical of the kernels of group homomorphisms, so we are led to the following definition.

**Definition 15.7.** Given any two groups $G$ and $G'$, a function $\varphi : G \to G'$ is a **homomorphism** iff

$$\varphi(g_1g_2) = \varphi(g_1)\varphi(g_2), \quad \text{for all } g_1, g_2 \in G.$$ 

Taking $g_1 = g_2 = e$ (in $G$), we see that

$$\varphi(e) = e',$$

and taking $g_1 = g$ and $g_2 = g^{-1}$, we see that

$$\varphi(g^{-1}) = (\varphi(g))^{-1}.$$
Example 15.4.

1. The map $\varphi : \mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$ given by $\varphi(m) = m \mod n$ for all $m \in \mathbb{Z}$ is a homomorphism.

2. The map $\text{det} : \text{GL}(n, \mathbb{R}) \to \mathbb{R}$ is a homomorphism because $\text{det}(AB) = \text{det}(A) \text{det}(B)$ for any two matrices $A, B$. Similarly, the map $\text{det} : \text{O}(n) \to \mathbb{R}$ is a homomorphism.

If $\varphi : G \to G'$ and $\psi : G' \to G''$ are group homomorphisms, then $\psi \circ \varphi : G \to G''$ is also a homomorphism. If $\varphi : G \to G'$ is a homomorphism of groups, and if $H \subseteq G$, $H' \subseteq G'$ are two subgroups, then it is easily checked that

$$\text{Im} \ H = \varphi(H) = \{ \varphi(g) \mid g \in H \}$$

is a subgroup of $G'$ and

$$\varphi^{-1}(H') = \{ g \in G \mid \varphi(g) \in H' \}$$

is a subgroup of $G$. In particular, when $H' = \{ e' \}$, we obtain the kernel, $\text{Ker} \ \varphi$, of $\varphi$.

Definition 15.8. If $\varphi : G \to G'$ is a homomorphism of groups, and if $H \subseteq G$ is a subgroup of $G$, then the subgroup of $G'$,

$$\text{Im} \ H = \varphi(H) = \{ \varphi(g) \mid g \in H \},$$

is called the image of $H$ by $\varphi$, and the subgroup of $G$,

$$\text{Ker} \ \varphi = \{ g \in G \mid \varphi(g) = e' \},$$

is called the kernel of $\varphi$.

Example 15.5.

1. The kernel of the homomorphism $\varphi : \mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$ is $n\mathbb{Z}$.

2. The kernel of the homomorphism $\text{det} : \text{GL}(n, \mathbb{R}) \to \mathbb{R}$ is $\text{SL}(n, \mathbb{R})$. Similarly, the kernel of the homomorphism $\text{det} : \text{O}(n) \to \mathbb{R}$ is $\text{SO}(n)$.

The following characterization of the injectivity of a group homomorphism is used all the time.

Proposition 15.8. If $\varphi : G \to G'$ is a homomorphism of groups, then $\varphi : G \to G'$ is injective iff $\text{Ker} \ \varphi = \{ e \}$. (We also write $\text{Ker} \ \varphi = (0)$.)

Proof. Assume $\varphi$ is injective. Since $\varphi(e) = e'$, if $\varphi(g) = e'$, then $\varphi(g) = \varphi(e)$, and by injectivity of $\varphi$ we must have $g = e$, so $\text{Ker} \ \varphi = \{ e \}$.

Conversely, assume that $\text{Ker} \ \varphi = \{ e \}$. If $\varphi(g_1) = \varphi(g_2)$, then by multiplication on the left by $(\varphi(g_1))^{-1}$ we get

$$e' = (\varphi(g_1))^{-1} \varphi(g_1) = (\varphi(g_1))^{-1} \varphi(g_2),$$

and so $\varphi(g_1) = \varphi(g_2)$, as desired.
and since \( \varphi \) is a homomorphism \((\varphi(g_1))^{-1} = \varphi(g_1^{-1})\), so
\[
e' = (\varphi(g_1))^{-1} \varphi(g_2) = \varphi(g_1^{-1}) \varphi(g_2) = \varphi(g_1^{-1}g_2).
\]
This shows that \( g_1^{-1}g_2 \in \text{Ker} \varphi \), but since \( \text{Ker} \varphi = \{e\} \) we have \( g_1^{-1}g_2 = e \), and thus \( g_2 = g_1 \), proving that \( \varphi \) is injective.

**Definition 15.9.** We say that a group homomorphism \( \varphi : G \to G' \) is an **isomorphism** if there is a homomorphism \( \psi : G' \to G \), so that
\[
\psi \circ \varphi = \text{id}_G \quad \text{and} \quad \varphi \circ \psi = \text{id}_{G'}.
\]
If \( \varphi \) is an isomorphism we say that the groups \( G \) and \( G' \) are **isomorphic**. When \( G' = G \), a group isomorphism is called an **automorphism**.

The reasoning used in the proof of Proposition 15.2 shows that if \( \varphi \) a group homomorphism \( \varphi : G \to G' \) is an isomorphism, then the homomorphism \( \psi : G' \to G \) satisfying Condition (†) is unique. This homomorphism is denoted \( \varphi^{-1} \).

The left translations \( L_g \) and the right translations \( R_g \) are automorphisms of \( G \).

Suppose \( \varphi : G \to G' \) is a bijective homomorphism, and let \( \varphi^{-1} \) be the inverse of \( \varphi \) (as a function). Then for all \( a, b \in G \), we have
\[
\varphi(\varphi^{-1}(a)\varphi^{-1}(b)) = \varphi(\varphi^{-1}(a))\varphi(\varphi^{-1}(b)) = ab,
\]
and so
\[
\varphi^{-1}(ab) = \varphi^{-1}(a)\varphi^{-1}(b),
\]
which proves that \( \varphi^{-1} \) is a homomorphism. Therefore, we proved the following fact.

**Proposition 15.9.** A bijective group homomorphism \( \varphi : G \to G' \) is an isomorphism.

Observe that the property
\[
gH = Hg, \quad \text{for all} \ g \in G.
\]
is equivalent by multiplication on the right by \( g^{-1} \) to
\[
gHg^{-1} = H, \quad \text{for all} \ g \in G,
\]
and the above is equivalent to
\[
gHg^{-1} \subseteq H, \quad \text{for all} \ g \in G.
\]
This is because \( gHg^{-1} \subseteq H \) implies \( H \subseteq g^{-1}Hg \), and this for all \( g \in G \).

**Proposition 15.10.** Let \( \varphi : G \to G' \) be a group homomorphism. Then \( H = \text{Ker} \varphi \) satisfies Property (**) and thus Property (**).
Proof. We have
\[ \varphi(ghg^{-1}) = \varphi(g)\varphi(h)\varphi(g^{-1}) = \varphi(g)e'\varphi(g)^{-1} = \varphi(g)\varphi(g)^{-1} = e', \]
for all \( h \in H = \text{Ker} \varphi \) and all \( g \in G \). Thus, by definition of \( H = \text{Ker} \varphi \), we have \( gHg^{-1} \subseteq H \). \( \square \)

**Definition 15.10.** For any group \( G \), a subgroup \( N \) of \( G \) is a *normal subgroup* of \( G \) iff
\[ gNg^{-1} = N, \quad \text{for all } g \in G. \]
This is denoted by \( N \triangleleft G \).

Proposition 15.10 shows that the kernel \( \text{Ker} \varphi \) of a homomorphism \( \varphi : G \rightarrow G' \) is a normal subgroup of \( G \).

Observe that if \( G \) is abelian, then *every* subgroup of \( G \) is normal.

Consider Example 15.2. Let \( R \in \text{SO}(2) \) and \( A \in \text{SL}(2, \mathbb{R}) \) be the matrices
\[ R = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}. \]
Then
\[ A^{-1} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \]
and we have
\[ ARA^{-1} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -2 \\ 1 & -1 \end{pmatrix}, \]
and clearly \( ARA^{-1} \notin \text{SO}(2) \). Therefore \( \text{SO}(2) \) is not a normal subgroup of \( \text{SL}(2, \mathbb{R}) \). The same counter-example shows that \( \text{O}(2) \) is not a normal subgroup of \( \text{GL}(2, \mathbb{R}) \).

Let \( R \in \text{SO}(2) \) and \( Q \in \text{SO}(3) \) be the matrices
\[ R = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad Q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}. \]
Then
\[ Q^{-1} = Q^\top = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix} \]
and we have
\[
QRQ^{-1} = \begin{pmatrix}
1 & 0 & 0 \\
0 & -1 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
0 & -1 & 0 \\
0 & 0 & -1 \\
1 & 0 & 0 \\
0 & 0 & 1 \\
1 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
1 & 0 & 0 \\
0 & 0 & 1 \\
0 & -1 & 0 \\
1 & 0 & 0 \\
0 & 0 & 1 \\
0 & -1 & 0 \\
1 & 0 & 0 \\
0 & 0 & 1 \\
1 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
1 & 0 & 0 \\
0 & 0 & 1 \\
0 & -1 & 0 \\
1 & 0 & 0 \\
0 & 0 & 1 \\
0 & -1 & 0 \\
1 & 0 & 0 \\
0 & 0 & 1 \\
1 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
1 & 0 & 0 \\
0 & 0 & 1 \\
0 & -1 & 0 \\
1 & 0 & 0 \\
0 & 0 & 1 \\
0 & -1 & 0 \\
1 & 0 & 0 \\
0 & 0 & 1 \\
1 & 0 & 0
\end{pmatrix}
= \begin{pmatrix}
0 & 0 & -1 \\
0 & 1 & 0 \\
1 & 0 & 0
\end{pmatrix}.
\]
Observe that $QRQ^{-1} \notin SO(2)$, so $SO(2)$ is not a normal subgroup of $SO(3)$.

Let $T$ and $A \in GL(2, \mathbb{R})$ be the following matrices
\[
T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.
\]
We have
\[
A^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = A,
\]
and
\[
ATA^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.
\]
The matrix $T$ is upper triangular, but $ATA^{-1}$ is not, so the group of $2 \times 2$ upper triangular matrices is not a normal subgroup of $GL(2, \mathbb{R})$.

Let $Q \in V$ and $A \in GL(2, \mathbb{R})$ be the following matrices
\[
Q = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.
\]
We have
\[
A^{-1} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}
\]
and
\[
AQA^{-1} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -2 \\ 0 & -1 \end{pmatrix}.
\]
Clearly $AQA^{-1} \notin V$, which shows that the Klein four group is not a normal subgroup of $GL(2, \mathbb{R})$.

The reader should check that the subgroups $n\mathbb{Z}$, $GL^+(n, \mathbb{R})$, $SL(n, \mathbb{R})$, and $SO(n, \mathbb{R})$ as a subgroup of $O(n, \mathbb{R})$, are normal subgroups.

If $N$ is a normal subgroup of $G$, the equivalence relation $\sim$ induced by left cosets (see Definition 15.5) is the same as the equivalence induced by right cosets. Furthermore, this equivalence relation is a congruence, which means that: For all $g_1, g_2, g'_1, g'_2 \in G$,
15.1. GROUPS, SUBGROUPS, COSETS

(1) If \( g_1N = g'_1N \) and \( g_2N = g'_2N \), then \( g_1g_2N = g'_1g'_2N \), and

(2) If \( g_1N = g_2N \), then \( g_1^{-1}N = g_2^{-1}N \).

As a consequence, we can define a group structure on the set \( G/ \sim \) of equivalence classes modulo \( \sim \), by setting

\[ (g_1N)(g_2N) = (g_1g_2)N. \]

**Definition 15.11.** Let \( G \) be a group and \( N \) be a normal subgroup of \( G \). The group obtained by defining the multiplication of (left) cosets by

\[ (g_1N)(g_2N) = (g_1g_2)N, \quad g_1, g_2 \in G \]

is denoted \( G/N \), and called the quotient of \( G \) by \( N \). The equivalence class \( gN \) of an element \( g \in G \) is also denoted \( [g] \). The map \( \pi: G \to G/N \) given by

\[ \pi(g) = \bar{g} = gN \]

is a group homomorphism called the canonical projection.

Since the kernel of a homomorphism is a normal subgroup, we obtain the following very useful result.

**Proposition 15.11.** Given a homomorphism of groups \( \varphi: G \to G' \), the groups \( G/\text{Ker } \varphi \) and \( \text{Im } \varphi = \varphi(G) \) are isomorphic.

**Proof.** Since \( \varphi \) is surjective onto its image, we may assume that \( \varphi \) is surjective, so that \( G' = \text{Im } \varphi \). We define a map \( \overline{\varphi}: G/\text{Ker } \varphi \to G' \) as follows:

\[ \overline{\varphi}(\bar{g}) = \varphi(g), \quad g \in G. \]

We need to check that the definition of this map does not depend on the representative chosen in the coset \( \bar{g} = g + \text{Ker } \varphi \), and that it is a homomorphism. If \( g' \) is another element in the coset \( g + \text{Ker } \varphi \), which means that \( g' = gh \) for some \( h \in \text{Ker } \varphi \), then

\[ \varphi(g') = \varphi(gh) = \varphi(g)\varphi(h) = \varphi(g)e' = \varphi(g), \]

since \( \varphi(h) = e' \) as \( h \in \text{Ker } \varphi \). This shows that

\[ \overline{\varphi}(\bar{g'}) = \overline{\varphi}(g') = \varphi(g) = \overline{\varphi}(\bar{g}), \]

so the map \( \overline{\varphi} \) is well defined. It is a homomorphism because

\[ \overline{\varphi}(\bar{gg'}) = \overline{\varphi}((gg')) = \varphi(gg') = \varphi(g)\varphi(g') = \overline{\varphi}(\bar{g})\overline{\varphi}(\bar{g'}). \]

The map \( \overline{\varphi} \) is injective because \( \overline{\varphi}(\bar{g}) = e' \) iff \( \varphi(g) = e' \) iff \( g \in \text{Ker } \varphi \), iff \( \bar{g} = \bar{e} \). The map \( \overline{\varphi} \) is surjective because \( \varphi \) is surjective. Therefore \( \overline{\varphi} \) is a bijective homomorphism, and thus an isomorphism, as claimed. \( \square \)
Proposition 15.11 is called the \emph{first isomorphism theorem}.

A useful way to construct groups is the \emph{direct product} construction.

\textbf{Definition 15.12.} Given two groups $G$ an $H$, we let $G \times H$ be the Cartesian product of the sets $G$ and $H$ with the multiplication operation $\cdot$ given by
\[
(g_1, h_1) \cdot (g_2, h_2) = (g_1 g_2, h_1 h_2).
\]
It is immediately verified that $G \times H$ is a group called the \emph{direct product} of $G$ and $H$.

Similarly, given any $n$ groups $G_1, \ldots, G_n$, we can define the direct product $G_1 \times \cdots \times G_n$ in a similar way.

If $G$ is an abelian group and $H_1, \ldots, H_n$ are subgroups of $G$, the situation is simpler. Consider the map
\[
a : H_1 \times \cdots \times H_n \to G
\]
given by
\[
a(h_1, \ldots, h_n) = h_1 + \cdots + h_n,
\]
using $+$ for the operation of the group $G$. It is easy to verify that $a$ is a group homomorphism, so its image is a subgroup of $G$ denoted by $H_1 + \cdots + H_n$, and called the \emph{sum} of the groups $H_i$. The following proposition will be needed.

\textbf{Proposition 15.12.} \textit{Given an abelian group $G$, if $H_1$ and $H_2$ are any subgroups of $G$ such that $H_1 \cap H_2 = \{0\}$, then the map $a$ is an isomorphism}
\[
a : H_1 \times H_2 \to H_1 + H_2.
\]
\textit{Proof.} The map is surjective by definition, so we just have to check that it is injective. For this, we show that $\ker a = \{(0, 0)\}$. We have $a(a_1, a_2) = 0$ iff $a_1 + a_2 = 0$ iff $a_1 = -a_2$. Since $a_1 \in H_1$ and $a_2 \in H_2$, we see that $a_1, a_2 \in H_1 \cap H_2 = \{0\}$, so $a_1 = a_2 = 0$, which proves that $\ker a = \{(0, 0)\}$. \hfill \Box

Under the conditions of Proposition 15.12, namely $H_1 \cap H_2 = \{0\}$, the group $H_1 + H_2$ is called the \emph{direct sum} of $H_1$ and $H_2$; it is denoted by $H_1 \oplus H_2$, and we have an isomorphism $H_1 \times H_2 \cong H_1 \oplus H_2$.

15.2 \textit{Group Actions: Part I, Definition and Examples}

If $X$ is a set (usually some kind of geometric space, for example, the sphere in $\mathbb{R}^3$, the upper half-plane, etc.), the “symmetries” of $X$ are often captured by the action of a group $G$ on $X$. In fact, if $G$ is a Lie group and the action satisfies some simple properties, the set $X$ can be given a manifold structure which makes it a projection (quotient) of $G$, a so-called “homogeneous space.”
Definition 15.13. Given a set $X$ and a group $G$, a left action of $G$ on $X$ (for short, an action of $G$ on $X$) is a function $\varphi: G \times X \to X$, such that:

1. For all $g, h \in G$ and all $x \in X$,
   \[ \varphi(g, \varphi(h, x)) = \varphi(gh, x), \]

2. For all $x \in X$,
   \[ \varphi(1, x) = x, \]
where $1 \in G$ is the identity element of $G$.

To alleviate the notation, we usually write $g \cdot x$ or even $gx$ for $\varphi(g, x)$, in which case the above axioms read:

1. For all $g, h \in G$ and all $x \in X$,
   \[ g \cdot (h \cdot x) = gh \cdot x, \]

2. For all $x \in X$,
   \[ 1 \cdot x = x. \]

The set $X$ is called a (left) $G$-set. The action $\varphi$ is faithful or effective iff for every $g$, if $g \cdot x = x$ for all $x \in X$, then $g = 1$. Faithful means that if the action of some element $g$ behaves like the identity, then $g$ must be the identity element. The action $\varphi$ is transitive iff for any two elements $x, y \in X$, there is some $g \in G$ so that $g \cdot x = y$.

Given an action $\varphi: G \times X \to X$, for every $g \in G$, we have a function $\varphi_g: X \to X$ defined by

\[ \varphi_g(x) = g \cdot x, \quad \text{for all } x \in X. \]

Observe that $\varphi_g$ has $\varphi_g^{-1}$ as inverse, since

\[ \varphi_g^{-1}(\varphi_g(x)) = \varphi_g^{-1}(g \cdot x) = g^{-1} \cdot (g \cdot x) = (g^{-1}g) \cdot x = 1 \cdot x = x, \]

and similarly, $\varphi_g \circ \varphi_g^{-1} = \text{id}$. Therefore, $\varphi_g$ is a bijection of $X$; that is, $\varphi_g$ is a permutation of $X$. Moreover, we check immediately that

\[ \varphi_g \circ \varphi_h = \varphi_{gh}, \]

so the map $g \mapsto \varphi_g$ is a group homomorphism from $G$ to $\mathfrak{S}_X$, the group of permutations of $X$. With a slight abuse of notation, this group homomorphism $G \longrightarrow \mathfrak{S}_X$ is also denoted $\varphi$.

Conversely, it is easy to see that any group homomorphism $\varphi: G \to \mathfrak{S}_X$ yields a group action $\cdot: G \times X \longrightarrow X$, by setting

\[ g \cdot x = \varphi(g)(x). \]

Observe that an action $\varphi$ is faithful iff the group homomorphism $\varphi: G \to \mathfrak{S}_X$ is injective, i.e. iff $\varphi$ has a trivial kernel. Also, we have $g \cdot x = y$ iff $g^{-1} \cdot y = x$, since $(gh) \cdot x = g \cdot (h \cdot x)$ and $1 \cdot x = x$, for all $g, h \in G$ and all $x \in X$. 
Definition 15.14. Given two $G$-sets $X$ and $Y$, a function $f: X \to Y$ is said to be equivariant, or a $G$-map, iff for all $x \in X$ and all $g \in G$, we have
\[ f(g \cdot x) = g \cdot f(x). \]
Equivalently, if the $G$-actions are denoted by $\varphi: G \times X \to X$ and $\psi: G \times Y \to Y$, we have the following commutative diagram for all $g \in G$:
\[
\begin{array}{ccc}
X & \xrightarrow{\varphi_g} & X \\
f \downarrow & & \downarrow f \\
Y & \xrightarrow{\psi_g} & Y.
\end{array}
\]
Remark: We can also define a right action $\cdot: X \times G \to X$ of a group $G$ on a set $X$ as a map satisfying the conditions
\begin{enumerate}
\item For all $g, h \in G$ and all $x \in X$,
\[ (x \cdot g) \cdot h = x \cdot gh, \]
\item For all $x \in X$,
\[ x \cdot 1 = x. \]
\end{enumerate}
Every notion defined for left actions is also defined for right actions in the obvious way.

However, one change is necessary. For every $g \in G$, the map $\varphi_g: X \to X$ must be defined as
\[ \varphi_g(x) = x \cdot g^{-1}, \]
in order for the map $g \mapsto \varphi_g$ from $G$ to $\mathfrak{S}_X$ to be a homomorphism $(\varphi_g \circ \varphi_h = \varphi_{gh})$.
Conversely, given a homomorphism $\varphi: G \to \mathfrak{S}_X$, we get a right action $\cdot: X \times G \to X$ by setting
\[ x \cdot g = \varphi(g^{-1})(x). \]
Here are some examples of (left) group actions.

Example 15.6. The unit sphere $S^2$ (more generally, $S^{n-1}$).
Recall that for any $n \geq 1$, the (real) unit sphere $S^{n-1}$ is the set of points in $\mathbb{R}^n$ given by
\[ S^{n-1} = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_1^2 + \cdots + x_n^2 = 1\}. \]
In particular, $S^2$ is the usual sphere in $\mathbb{R}^3$. Since the group $\text{SO}(3) = \text{SO}(3, \mathbb{R})$ consists of (orientation preserving) linear isometries, i.e., linear maps that are distance preserving (and of determinant $+1$), and every linear map leaves the origin fixed, we see that any rotation maps $S^2$ into itself.
Beware that this would be false if we considered the group of affine isometries $\text{SE}(3)$ of $\mathbb{E}^3$. For example, a screw motion does not map $S^2$ into itself, even though it is distance preserving, because the origin is translated.

Thus, for $X = S^2$ and $G = \text{SO}(3)$, we have an action $\cdot: \text{SO}(3) \times S^2 \to S^2$, given by the matrix multiplication

$$R \cdot x = Rx.$$ 

The verification that the above is indeed an action is trivial. This action is transitive. This is because, for any two points $x, y$ on the sphere $S^2$, there is a rotation whose axis is perpendicular to the plane containing $x, y$ and the center $O$ of the sphere (this plane is not unique when $x$ and $y$ are antipodal, i.e., on a diameter) mapping $x$ to $y$. See Figure 15.1.

![Figure 15.1: The rotation which maps $x$ to $y$.](image)

Similarly, for any $n \geq 1$, let $X = S^{n-1}$ and $G = \text{SO}(n)$ and define the action $\cdot: \text{SO}(n) \times S^{n-1} \to S^{n-1}$ as $R \cdot x = Rx$. It is easy to show that this action is transitive.

Analogously, we can define the (complex) unit sphere $\Sigma^{n-1}$, as the set of points in $\mathbb{C}^n$ given by

$$\Sigma^{n-1} = \{ (z_1, \ldots, z_n) \in \mathbb{C}^n \mid z_1 \bar{z}_1 + \cdots + z_n \bar{z}_n = 1 \}.$$ 

If we write $z_j = x_j + iy_j$, with $x_j, y_j \in \mathbb{R}$, then

$$\Sigma^{n-1} = \{ (x_1, \ldots, x_n, y_1, \ldots, y_n) \in \mathbb{R}^{2n} \mid x_1^2 + \cdots + x_n^2 + y_1^2 + \cdots + y_n^2 = 1 \}.$$ 

Therefore, we can view the complex sphere $\Sigma^{n-1}$ (in $\mathbb{C}^n$) as the real sphere $S^{2n-1}$ (in $\mathbb{R}^{2n}$). By analogy with the real case, we can define for $X = \Sigma^{n-1}$ and $G = \text{SU}(n)$ an action $\cdot: \text{SU}(n) \times \Sigma^{n-1} \to \Sigma^{n-1}$ of the group $\text{SU}(n)$ of linear maps of $\mathbb{C}^n$ preserving the Hermitian inner product (and the origin, as all linear maps do), and this action is transitive.
One should not confuse the unit sphere $\Sigma^{n-1}$ with the hypersurface $S^{n-1}_C$, given by

$$S^{n-1}_C = \{(z_1, \ldots, z_n) \in C^n \mid z_1^2 + \cdots + z_n^2 = 1\}.$$ 

For instance, one should check that a line $L$ through the origin intersects $\Sigma^{n-1}$ in a circle, whereas it intersects $S^{n-1}_C$ in exactly two points! Recall for a fixed $u = (x_1, \ldots, x_n, y_1, \ldots, y_n) \in C^n$, that $L = \{\gamma u \mid \gamma \in C\}$. Since $\gamma = \rho (\cos \theta + i \sin \theta)$, we deduce that $L$ is actually the two dimensional subspace through the origin spanned by the orthogonal vectors $(x_1, \ldots, x_n, y_1, \ldots, y_n)$ and $(-y_1, \ldots, -y_n, x_1, \ldots, x_n)$.

**Example 15.7.** The upper half-plane.

The upper half-plane $H$ is the open subset of $\mathbb{R}^2$ consisting of all points $(x, y) \in \mathbb{R}^2$, with $y > 0$. It is convenient to identify $H$ with the set of complex numbers $z \in C$ such that $\Im(z) > 0$. Then we can let $X = H$ and $G = \text{SL}(2, \mathbb{R})$ and define an action $\cdot : \text{SL}(2, \mathbb{R}) \times H \to H$ of the group $\text{SL}(2, \mathbb{R})$ on $H$, as follows: For any $z \in H$, for any $A \in \text{SL}(2, \mathbb{R})$,

$$A \cdot z = \frac{az + b}{cz + d},$$

where

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

with $ad - bc = 1$.

It is easily verified that $A \cdot z$ is indeed always well defined and in $H$ when $z \in H$ (check this). To see why this action is transitive, let $z$ and $w$ be two arbitrary points of $H$ where $z = x + iy$ and $w = u + iv$ with $x, u \in \mathbb{R}$ and $y, v \in \mathbb{R}^+$ (i.e. $y$ and $v$ are positive real numbers). Define $A = \begin{pmatrix} \sqrt{\frac{u}{y}} & \frac{uy-vz}{\sqrt{uv}} \\ 0 & \sqrt{\frac{v}{u}} \end{pmatrix}$. Note that $A \in \text{SL}(2, \mathbb{R})$. A routine calculation shows that $A \cdot z = w$.

Before introducing Example 15.8, we need to define the groups of Möbius transformations and the Riemann sphere. Maps of the form

$$z \mapsto \frac{az + b}{cz + d},$$

where $z \in C$ and $ad - bc = 1$, are called Möbius transformations. Here, $a, b, c, d \in \mathbb{R}$, but in general, we allow $a, b, c, d \in C$. Actually, these transformations are not necessarily defined everywhere on $C$, for example, for $z = -d/c$ if $c \neq 0$. To fix this problem, we add a “point at infinity” $\infty$ to $C$, and define Möbius transformations as functions $C \cup \{\infty\} \to C \cup \{\infty\}$. If $c = 0$, the Möbius transformation sends $\infty$ to itself, otherwise, $-d/c \mapsto \infty$ and $\infty \mapsto a/c$.

The space $C \cup \{\infty\}$ can be viewed as the plane $\mathbb{R}^2$ extended with a point at infinity. Using a stereographic projection from the sphere $S^2$ to the plane (say from the north pole to the
equatorial plane), we see that there is a bijection between the sphere $S^2$ and $\mathbb{C} \cup \{\infty\}$. More precisely, the stereographic projection $\sigma_N$ of the sphere $S^2$ from the north pole $N = (0, 0, 1)$ to the plane $z = 0$ (extended with the point at infinity $\infty$) is given by

$$(x, y, z) \in S^2 - \{(0, 0, 1)\} \mapsto \left(\frac{x}{1-z}, \frac{y}{1-z}\right) = \frac{x + iy}{1-z} \in \mathbb{C}, \quad \text{with} \quad (0, 0, 1) \mapsto \infty.$$ 

The inverse stereographic projection $\sigma_N^{-1}$ is given by

$$(u, v) \mapsto \left(\frac{2u}{u^2 + v^2 + 1}, \frac{2v}{u^2 + v^2 + 1}, \frac{u^2 + v^2 - 1}{u^2 + v^2 + 1}\right), \quad \text{with} \quad \infty \mapsto (0, 0, 1).$$ 

Intuitively, the inverse stereographic projection “wraps” the equatorial plane around the sphere. See Figure 15.2.

Figure 15.2: Inverse stereographic projection.

The space $\mathbb{C} \cup \{\infty\}$ is known as the Riemann sphere. We will see shortly that $\mathbb{C} \cup \{\infty\} \cong S^2$ is also the complex projective line $\mathbb{CP}^1$. In summary, Möbius transformations are bijections of the Riemann sphere. It is easy to check that these transformations form a group under composition for all $a, b, c, d \in \mathbb{C}$, with $ad - bc = 1$. This is the Möbius group, denoted $\text{M"{o}b}^+$. The Möbius transformations corresponding to the case $a, b, c, d \in \mathbb{R}$, with $ad - bc = 1$ form a subgroup of $\text{M"{o}b}^+$ denoted $\text{M"{o}b}_R^+$.

The map from $\text{SL}(2, \mathbb{C})$ to $\text{M"{o}b}^+$ that sends $A \in \text{SL}(2, \mathbb{C})$ to the corresponding Möbius transformation is a surjective group homomorphism, and one checks easily that its kernel is $\{-I, I\}$ (where $I$ is the $2 \times 2$ identity matrix). Therefore, the Möbius group $\text{M"{o}b}^+$ is
isomorphic to the quotient group $\text{SL}(2, \mathbb{C})/\{-I, I\}$, denoted $\text{PSL}(2, \mathbb{C})$. This latter group turns out to be the group of projective transformations of the projective space $\mathbb{CP}^1$. The same reasoning shows that the subgroup $\text{M"ob}^+_\mathbb{R}$ is isomorphic to $\text{SL}(2, \mathbb{R})/\{-I, I\}$, denoted $\text{PSL}(2, \mathbb{R})$.

**Example 15.8.** The Riemann sphere $\mathbb{C} \cup \{\infty\}$.

Let $X = \mathbb{C} \cup \{\infty\}$ and $G = \text{SL}(2, \mathbb{C})$. The group $\text{SL}(2, \mathbb{C})$ acts on $\mathbb{C} \cup \{\infty\} \cong S^2$ the same way that $\text{SL}(2, \mathbb{R})$ acts on $H$, namely: For any $A \in \text{SL}(2, \mathbb{C})$, for any $z \in \mathbb{C} \cup \{\infty\}$,

$$A \cdot z = \frac{az + b}{cz + d},$$

where

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{with} \quad ad - bc = 1.$$  

This action is transitive, an exercise we leave for the reader.

**Example 15.9.** The unit disk.

One may recall from complex analysis that the scaled (complex) M"obius transformation

$$z \mapsto \frac{z - i}{z + i}$$

is a biholomorphic or analytic isomorphism between the upper half plane $H$ and the open unit disk

$$D = \{z \in \mathbb{C} \mid |z| < 1\}.$$  

As a consequence, it is possible to define a transitive action of $\text{SL}(2, \mathbb{R})$ on $D$. This can be done in a more direct fashion, using a group isomorphic to $\text{SL}(2, \mathbb{R})$, namely, $\text{SU}(1, 1)$ (a group of complex matrices), but we don’t want to do this right now.

**Example 15.10.** The unit Riemann sphere revisited.

Another interesting action is the action of $\text{SU}(2)$ on the extended plane $\mathbb{C} \cup \{\infty\}$. Recall that the group $\text{SU}(2)$ consists of all complex matrices of the form

$$A = \begin{pmatrix} \alpha & \beta \\ -\bar{\beta} & \bar{\alpha} \end{pmatrix} \quad \alpha, \beta \in \mathbb{C}, \quad \alpha \bar{\alpha} + \beta \bar{\beta} = 1,$$

Let $X = \mathbb{C} \cup \{\infty\}$ and $G = \text{SU}(2)$. The action $\cdot : \text{SU}(2) \times (\mathbb{C} \cup \{\infty\}) \to \mathbb{C} \cup \{\infty\}$ is given by

$$A \cdot w = \frac{\alpha w + \beta}{-\bar{\beta} w + \bar{\alpha}}, \quad w \in \mathbb{C} \cup \{\infty\}.$$  

This action is transitive, but the proof of this fact relies on the surjectivity of the group homomorphism

$$\rho : \text{SU}(2) \to \text{SO}(3)$$
defined below, and the stereographic projection $\sigma_N$ from $S^2$ onto $\mathbb{C} \cup \{\infty\}$. In particular, take $z, w \in \mathbb{C} \cup \{\infty\}$, use the inverse stereographic projection to obtain two points on $S^2$, namely $\sigma_N^{-1}(z)$ and $\sigma_N^{-1}(w)$. Then apply the appropriate rotation $R \in \text{SO}(3)$ to map $\sigma_N^{-1}(z)$ onto $\sigma_N^{-1}(w)$. Such a rotation exists by the argument presented in Example 15.6. Since $\rho: \text{SU}(2) \to \text{SO}(3)$ is surjective (see below), we know there must exist $A \in \text{SU}(2)$ such that $\rho(A) = R$ and $A \cdot z = w$.

Using the stereographic projection $\sigma_N$ from $S^2$ onto $\mathbb{C} \cup \{\infty\}$ and its inverse $\sigma_N^{-1}$, we can define an action of $\text{SU}(2)$ on $S^2$ by

$$A \cdot (x, y, z) = \sigma_N^{-1}(A \cdot \sigma_N(x, y, z)), \quad (x, y, z) \in S^2.$$ 

Although this is not immediately obvious, it turns out that $\text{SU}(2)$ acts on $S^2$ by maps that are restrictions of linear maps to $S^2$, and since these linear maps preserve $S^2$, they are orthogonal transformations. Thus, we obtain a continuous (in fact, smooth) group homomorphism

$$\rho: \text{SU}(2) \to \text{O}(3).$$

Since $\text{SU}(2)$ is connected and $\rho$ is continuous, the image of $\text{SU}(2)$ is contained in the connected component of $I$ in $\text{O}(3)$, namely $\text{SO}(3)$, so $\rho$ is a homomorphism

$$\rho: \text{SU}(2) \to \text{SO}(3).$$

We will see that this homomorphism is surjective and that its kernel is $\{I, -I\}$. The upshot is that we have an isomorphism

$$\text{SO}(3) \cong \text{SU}(2)/\{I, -I\}.$$ 

The homomorphism $\rho$ is a way of describing how a unit quaternion (any element of $\text{SU}(2)$) induces a rotation, via the stereographic projection and its inverse. If we write $\alpha = a + ib$ and $\beta = c + id$, a rather tedious computation yields

$$\rho(A) = \begin{pmatrix} a^2 - b^2 - c^2 + d^2 & -2ab - 2cd & -2ac + 2bd \\ 2ab - 2cd & a^2 - b^2 + c^2 - d^2 & -2ad - 2bc \\ 2ac + 2bd & 2ad - 2bc & a^2 + b^2 - c^2 - d^2 \end{pmatrix}.$$ 

One can check that $\rho(A)$ is indeed a rotation matrix which represents the rotation whose axis is the line determined by the vector $(d, -c, b)$ and whose angle $\theta \in [-\pi, \pi]$ is determined by

$$\cos \frac{\theta}{2} = |a|.$$ 

We can also compute the derivative $d\rho_I: \mathfrak{su}(2) \to \mathfrak{so}(3)$ of $\rho$ at $I$ as follows. Recall that $\mathfrak{su}(2)$ consists of all complex matrices of the form

$$\begin{pmatrix} ib & c + id \\ -c + id & -ib \end{pmatrix}, \quad b, c, d \in \mathbb{R},$$
so pick the following basis for $\mathfrak{su}(2)$,

\[
X_1 = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \quad X_2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad X_3 = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix},
\]

and define the curves in $\mathbf{SU}(2)$ through $I$ given by

\[
c_1(t) = \begin{pmatrix} e^{it} & 0 \\ 0 & e^{-it} \end{pmatrix}, \quad c_2(t) = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}, \quad c_3(t) = \begin{pmatrix} \cos t & i \sin t \\ i \sin t & \cos t \end{pmatrix}.
\]

It is easy to check that $c_i'(0) = X_i$ for $i = 1, 2, 3$, and that

\[
d\rho_I(X_1) = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad d\rho_I(X_2) = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad d\rho_I(X_3) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}.
\]

Thus we have

\[
d\rho_I(X_1) = 2E_3, \quad d\rho_I(X_2) = -2E_2, \quad d\rho_I(X_3) = 2E_1,
\]

where $(E_1, E_2, E_3)$ is the basis of $\mathfrak{so}(3)$ given by

\[
E_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \quad E_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \quad E_3 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},
\]

which means that $d\rho_I$ is an isomorphism between the Lie algebras $\mathfrak{su}(2)$ and $\mathfrak{so}(3)$.

Recall that we have the commutative diagram

\[
\begin{array}{ccc}
\mathbf{SU}(2) & \overset{\rho}{\longrightarrow} & \mathbf{SO}(3) \\
\downarrow{\text{exp}} & & \downarrow{\text{exp}} \\
\mathfrak{su}(2) & \overset{d\rho_I}{\longrightarrow} & \mathfrak{so}(3).
\end{array}
\]

Since $d\rho_I$ is surjective and the exponential map $\exp: \mathfrak{so}(3) \to \mathbf{SO}(3)$ is surjective, we conclude that $\rho$ is surjective. (We also know that $\exp: \mathfrak{su}(2) \to \mathbf{SU}(2)$ is surjective.) Observe that $\rho(-A) = \rho(A)$, and it is easy to check that $\text{Ker} \rho = \{I, -I\}$.

**Example 15.11.** The set of $n \times n$ symmetric, positive, definite matrices, $\text{SPD}(n)$.

Let $X = \text{SPD}(n)$ and $G = \mathbf{GL}(n)$. The group $\mathbf{GL}(n) = \mathbf{GL}(n, \mathbb{R})$ acts on $\text{SPD}(n)$ as follows: for all $A \in \mathbf{GL}(n)$ and all $S \in \text{SPD}(n)$,

\[
A \cdot S = ASA^\top.
\]

It is easily checked that $ASA^\top$ is in $\text{SPD}(n)$ if $S$ is in $\text{SPD}(n)$. First observe that $ASA^\top$ is symmetric since

\[
(ASA^\top)^\top = AS^\top A^\top = ASA^\top.
\]
Next recall the following characterization of positive definite matrix, namely
\[ y^\top S y > 0, \quad \text{whenever } y \neq 0. \]
We want to show \( x^\top (A^\top S A) x > 0 \) for all \( x \neq 0 \). Since \( A \) is invertible, we have \( x = A^{-1} y \) for some nonzero \( y \), and hence
\[
x^\top (A^\top S A) x = y^\top (A^{-1})^\top A^\top S A^{-1} y = y^\top S y > 0.
\]
Hence \( A^\top S A \) is positive definite. This action is transitive because every SPD matrix \( S \) can be written as \( S = A A^\top \), for some invertible matrix \( A \) (prove this as an exercise). Given any two SPD matrices \( S_1 = A_1 A_1^\top \) and \( S_2 = A_2 A_2^\top \) with \( A_1 \) and \( A_2 \) invertible, if \( A = A_2 A_1^{-1} \), we have
\[
A \cdot S_1 = A_2 A_1^{-1} S_1 (A_2 A_1^{-1})^\top = A_2 A_1^{-1} S_1 (A_1^\top)^{-1} A_2^\top = A_2 A_1^{-1} A_1 A_1^\top (A_1^\top)^{-1} A_2^\top = A_2 A_2^\top = S_2.
\]

**Example 15.12.** The projective spaces \( \mathbb{R}P^n \) and \( \mathbb{C}P^n \).

The *(real)* projective space \( \mathbb{R}P^n \) is the set of all lines through the origin in \( \mathbb{R}^{n+1} \); that is, the set of one-dimensional subspaces of \( \mathbb{R}^{n+1} \) (where \( n \geq 0 \)). Since a one-dimensional subspace \( L \subseteq \mathbb{R}^{n+1} \) is spanned by any nonzero vector \( u \in L \), we can view \( \mathbb{R}P^n \) as the set of equivalence classes of nonzero vectors in \( \mathbb{R}^{n+1} - \{0\} \) modulo the equivalence relation
\[
u \sim v \quad \text{iff} \quad v = \lambda u, \quad \text{for some } \lambda \in \mathbb{R}, \lambda \neq 0.
\]
In terms of this definition, there is a projection \( \text{pr}: (\mathbb{R}^{n+1} - \{0\}) \to \mathbb{R}P^n \), given by \( \text{pr}(u) = [u]_{\sim} \), the equivalence class of \( u \) modulo \( \sim \). Write \([u]\) for the line defined by the nonzero vector \( u \). Since every line \( L \) in \( \mathbb{R}^{n+1} \) intersects the sphere \( S^n \) in two antipodal points, we can view \( \mathbb{R}P^n \) as the quotient of the sphere \( S^n \) by identification of antipodal points. See Figures 15.3 and 15.4.

Let \( X = \mathbb{R}P^n \) and \( G = \text{SO}(n + 1) \). We define an action of \( \text{SO}(n + 1) \) on \( \mathbb{R}P^n \) as follows: For any line \( L = [u] \), for any \( R \in \text{SO}(n + 1) \),
\[
R \cdot L = [Ru].
\]
Since \( R \) is linear, the line \([Ru]\) is well defined; that is, does not depend on the choice of \( u \in L \). The reader can show that this action is transitive.

The *(complex)* projective space \( \mathbb{C}P^n \) is defined analogously as the set of all lines through the origin in \( \mathbb{C}^{n+1} \); that is, the set of one-dimensional subspaces of \( \mathbb{C}^{n+1} \) (where \( n \geq 0 \)). This time, we can view \( \mathbb{C}P^n \) as the set of equivalence classes of vectors in \( \mathbb{C}^{n+1} - \{0\} \) modulo the equivalence relation
\[
u \sim v \quad \text{iff} \quad v = \lambda u, \quad \text{for some } \lambda \neq 0 \in \mathbb{C}.
\]
Figure 15.3: Three constructions for $\mathbb{RP}^1 \cong S^1$. Illustration (i.) applies the equivalence relation. Since any line through the origin, excluding the $x$-axis, intersects the line $y = 1$, its equivalence class is represented by its point of intersection on $y = 1$. Hence, $\mathbb{RP}^n$ is the disjoint union of the line $y = 1$ and the point of infinity given by the $x$-axis. Illustration (ii.) represents $\mathbb{RP}^1$ as the quotient of the circle $S^1$ by identification of antipodal points. Illustration (iii.) is a variation which glues the equatorial points of the upper semicircle.

We have the projection $pr: \mathbb{C}^{n+1} - \{0\} \to \mathbb{CP}^n$, given by $pr(u) = [u]$, the equivalence class of $u$ modulo $\sim$. Again, write $[u]$ for the line defined by the nonzero vector $u$. Let $X = \mathbb{CP}^n$ and $G = SU(n + 1)$. We define an action of $SU(n + 1)$ on $\mathbb{CP}^n$ as follows: For any line $L = [u]$, for any $R \in SU(n + 1)$,

$$R \cdot L = [Ru].$$

Again, this action is well defined and it is transitive. (Check this.)

Before progressing to our final example of group actions, we take a moment to construct $\mathbb{CP}^n$ as a quotient space of $S^{2n+1}$. Recall that $\Sigma^n \subseteq \mathbb{C}^{n+1}$, the unit sphere in $\mathbb{C}^{n+1}$, is defined by

$$\Sigma^n = \{(z_1, \ldots, z_{n+1}) \in \mathbb{C}^{n+1} \mid z_1 \bar{z}_1 + \cdots + z_{n+1} \bar{z}_{n+1} = 1\}.$$

For any line $L = [u]$, where $u \in \mathbb{C}^{n+1}$ is a nonzero vector, writing $u = (u_1, \ldots, u_{n+1})$, a point $z \in \mathbb{C}^{n+1}$ belongs to $L$ iff $z = \lambda(u_1, \ldots, u_{n+1})$, for some $\lambda \in \mathbb{C}$. Therefore, the intersection
15.2. GROUP ACTIONS: PART I, DEFINITION AND EXAMPLES

Figure 15.4: Three constructions for $\mathbb{R}P^2$. Illustration (i.) applies the equivalence relation. Since any line through the origin which is not contained in the $xy$-plane intersects the plane $z = 1$, its equivalence class is represented by its point of intersection on $z = 1$. Hence, $\mathbb{R}P^2$ is the disjoint union of the plane $z = 1$ and the copy of $\mathbb{R}P^1$ provided by the $xy$-plane. Illustration (ii.) represents $\mathbb{R}P^2$ as the quotient of the sphere $S^2$ by identification of antipodal points. Illustration (iii.) is a variation which glues the antipodal points on boundary of the unit disk, which is represented as as the upper hemisphere.

$L \cap \Sigma^n$ of the line $L$ and the sphere $\Sigma^n$ is given by

$$L \cap \Sigma^n = \{\lambda(u_1, \ldots, u_{n+1}) \in \mathbb{C}^{n+1} \mid \lambda \in \mathbb{C}, \lambda \bar{\lambda}(u_1\bar{u}_1 + \cdots + u_{n+1}\bar{u}_{n+1}) = 1\},$$
i.e.,

$$L \cap \Sigma^n = \left\{\lambda(u_1, \ldots, u_{n+1}) \in \mathbb{C}^{n+1} \mid \lambda \in \mathbb{C}, |\lambda| = \frac{1}{\sqrt{|u_1|^2 + \cdots + |u_{n+1}|^2}}\right\}.$$  

Thus, we see that there is a bijection between $L \cap \Sigma^n$ and the circle $S^1$; that is, geometrically $L \cap \Sigma^n$ is a circle. Moreover, since any line $L$ through the origin is determined by just one other point, we see that for any two lines $L_1$ and $L_2$ through the origin,

$$L_1 \neq L_2 \iff (L_1 \cap \Sigma^n) \cap (L_2 \cap \Sigma^n) = \emptyset.$$  

However, $\Sigma^n$ is the sphere $S^{2n+1}$ in $\mathbb{R}^{2n+2}$. It follows that $\mathbb{CP}^n$ is the quotient of $S^{2n+1}$ by the equivalence relation $\sim$ defined such that

$$y \sim z \iff y, z \in L \cap \Sigma^n, \text{ for some line, } L, \text{ through the origin.}$$
Therefore, we can write

\[ S^{2n+1}/S^1 \cong \mathbb{CP}^n. \]

The case \( n = 1 \) is particularly interesting, as it turns out that

\[ S^3/S^1 \cong S^2. \]

This is the famous Hopf fibration. To show this, proceed as follows: As

\[ S^3 \cong \Sigma^1 = \{(z, z') \in \mathbb{C}^2 \mid |z|^2 + |z'|^2 = 1\}, \]

define a map, \( HF: S^3 \to S^2 \), by

\[ HF((z, z')) = (2z\overline{z'}, |z|^2 - |z'|^2). \]

We leave as a homework exercise to prove that this map has range \( S^2 \) and that

\[ HF((z_1, z'_1)) = HF((z_2, z'_2)) \iff (z_1, z'_1) = \lambda(z_2, z'_2), \quad \text{for some } \lambda \text{ with } |\lambda| = 1. \]

In other words, for any point, \( p \in S^2 \), the inverse image \( HF^{-1}(p) \) (also called fibre over \( p \)) is a circle on \( S^3 \). Consequently, \( S^3 \) can be viewed as the union of a family of disjoint circles. This is the Hopf fibration. It is possible to visualize the Hopf fibration using the stereographic projection from \( S^3 \) onto \( \mathbb{R}^3 \). This is a beautiful and puzzling picture. For example, see Berger [2]. Therefore, \( HF \) induces a bijection from \( \mathbb{CP}^1 \) to \( S^2 \), and it is a homeomorphism.

**Example 15.13.** Affine spaces.

Let \( X \) be a set and \( E \) a real vector space. A transitive and faithful action \( \cdot : E \times X \to X \) of the additive group of \( E \) on \( X \) makes \( X \) into an affine space. The intuition is that the members of \( E \) are translations.

Those familiar with affine spaces as in Gallier [20] (Chapter 2) or Berger [2] will point out that if \( X \) is an affine space, then not only is the action of \( E \) on \( X \) transitive, but more is true: For any two points \( a, b \in E \), there is a unique vector \( u \in E \), such that \( u \cdot a = b \). By the way, the action of \( E \) on \( X \) is usually considered to be a right action and is written additively, so \( u \cdot a \) is written \( a + u \) (the result of translating \( a \) by \( u \)). Thus, it would seem that we have to require more of our action. However, this is not necessary because \( E \) (under addition) is abelian. More precisely, we have the proposition

**Proposition 15.13.** If \( G \) is an abelian group acting on a set \( X \) and the action \( \cdot : G \times X \to X \) is transitive and faithful, then for any two elements \( x, y \in X \), there is a unique \( g \in G \) so that \( g \cdot x = y \) (the action is simply transitive).

**Proof.** Since our action is transitive, there is at least some \( g \in G \) so that \( g \cdot x = y \). Assume that we have \( g_1, g_2 \in G \) with

\[ g_1 \cdot x = g_2 \cdot x = y. \]
15.3 GROUP ACTIONS: PART II, STABILIZERS AND HOMOGENEOUS SPACES

We shall prove that
\[ g_1 \cdot z = g_2 \cdot z, \quad \text{for all } z \in X. \]
This implies that
\[ g_1 g_2^{-1} \cdot z = z, \quad \text{for all } z \in X. \]
As our action is faithful, \( g_1 g_2^{-1} = 1 \), and we must have \( g_1 = g_2 \), which proves our proposition.

Pick any \( z \in X \). As our action is transitive, there is some \( h \in G \) so that \( z = h \cdot x \). Then, we have
\[
g_1 \cdot z &= g_1 \cdot (h \cdot x) \\
&= (g_1 h) \cdot x \\
&= (hg_1) \cdot x \quad \text{(since } G \text{ is abelian)} \\
&= h \cdot (g_1 \cdot x) \\
&= h \cdot (g_2 \cdot x) \quad \text{(since } g_1 \cdot x = g_2 \cdot x) \\
&= (hg_2) \cdot x \\
&= (g_2 h) \cdot x \quad \text{(since } G \text{ is abelian)} \\
&= g_2 \cdot (h \cdot x) \\
&= g_2 \cdot z.
\]
Therefore, \( g_1 \cdot z = g_2 \cdot z \) for all \( z \in X \), as claimed. \( \square \)

15.3 Group Actions: Part II, Stabilizers and Homogeneous Spaces

Now that we have an understanding of how a group \( G \) acts on a set \( X \), we may use this action to form new topological spaces, namely homogeneous spaces. In the construction of homogeneous spaces, the subset of group elements that leaves some given element \( x \in X \) fixed plays an important role.

Definition 15.15. Given an action \( \cdot : G \times X \to X \) of a group \( G \) on a set \( X \), for any \( x \in X \), the group \( G_x \) (also denoted \( \text{Stab}_G(x) \)), called the stabilizer of \( x \) or isotropy group at \( x \), is given by
\[ G_x = \{ g \in G \mid g \cdot x = x \}. \]

We have to verify that \( G_x \) is indeed a subgroup of \( G \), but this is easy. Indeed, if \( g \cdot x = x \) and \( h \cdot x = x \), then we also have \( h^{-1} \cdot x = x \) and so, we get \( gh^{-1} \cdot x = x \), proving that \( G_x \) is a subgroup of \( G \). In general, \( G_x \) is not a normal subgroup.
Observe that
\[ G_{g \cdot x} = gG_xg^{-1}, \]
for all \( g \in G \) and all \( x \in X \). Indeed,
\[
G_{g \cdot x} = \{ h \in G \mid h \cdot (g \cdot x) = g \cdot x \} = \{ h \in G \mid hg \cdot x = g \cdot x \} = \{ h \in G \mid g^{-1}hg \cdot x = x \},
\]
which shows \( g^{-1}G_{g \cdot x}g \subseteq G_x \), or equivalently that \( G_{g \cdot x} \subseteq gG_xg^{-1} \). It remains to show that \( gG_xg^{-1} \subseteq G_{g \cdot x} \). Take an element of \( gG_xg^{-1} \), which has the form \( ghg^{-1} \) with \( h \cdot x = x \). Since \( h \cdot x = x \), we have \((ghg^{-1}) \cdot gx = gx\), which shows that \( ghg^{-1} \in G_{g \cdot x} \).

Because \( G_{g \cdot x} = gG_xg^{-1} \), the stabilizers of \( x \) and \( g \cdot x \) are conjugate of each other.

When the action of \( G \) on \( X \) is transitive, for any fixed \( x \in G \), the set \( X \) is a quotient (as a set, not as group) of \( G \) by \( G_x \). Indeed, we can define the map, \( \pi_x : G \to X \), by
\[ \pi_x(g) = g \cdot x, \quad \text{for all } g \in G. \]
Observe that
\[ \pi_x(gG_x) = (gG_x) \cdot x = g \cdot (G_x \cdot x) = g \cdot x = \pi_x(g). \]
This shows that \( \pi_x : G \to X \) induces a quotient map \( \pi_x : G/G_x \to X \), from the set \( G/G_x \) of (left) cosets of \( G_x \) to \( X \), defined by
\[ \pi_x(gG_x) = g \cdot x. \]
Since
\[ \pi_x(g) = \pi_x(h) \quad \text{iff} \quad g \cdot x = h \cdot x \quad \text{iff} \quad g^{-1}h \cdot x = x \quad \text{iff} \quad g^{-1}h \in G_x \quad \text{iff} \quad gG_x = hG_x, \]
we deduce that \( \pi_x : G/G_x \to X \) is injective. However, since our action is transitive, for every \( y \in X \), there is some \( g \in G \) so that \( g \cdot x = y \), and so \( \pi_x(gG_x) = g \cdot x = y \); that is, the map \( \pi_x \) is also surjective. Therefore, the map \( \pi_x : G/G_x \to X \) is a bijection (of sets, not groups). The map \( \pi_x : G \to X \) is also surjective. Let us record this important fact as

**Proposition 15.14.** If \( \cdot : G \times X \to X \) is a transitive action of a group \( G \) on a set \( X \), for every fixed \( x \in X \), the surjection \( \pi_x : G \to X \) given by
\[ \pi_x(g) = g \cdot x \]
duces a bijection
\[ \pi_x : G/G_x \to X, \]
where \( G_x \) is the stabilizer of \( x \). See Figure 15.5.
Figure 15.5: A schematic representation of $G/G_x \cong X$, where $G$ is the gray solid, $X$ is its purple circular base, and $G_x$ is the pink vertical strand. The dotted strands are the fibres $gG_x$.

The map $\pi_x: G \to X$ (corresponding to a fixed $x \in X$) is sometimes called a projection of $G$ onto $X$. Proposition 15.14 shows that for every $y \in X$, the subset $\pi_x^{-1}(y)$ of $G$ (called the fibre above $y$) is equal to some coset $gG_x$ of $G$, and thus is in bijection with the group $G_x$ itself. We can think of $G$ as a moving family of fibres $G_x$ parametrized by $X$. This point of view of viewing a space as a moving family of simpler spaces is typical in (algebraic) geometry, and underlies the notion of (principal) fibre bundle.

Note that if the action $\cdot: G \times X \to X$ is transitive, then the stabilizers $G_x$ and $G_y$ of any two elements $x, y \in X$ are isomorphic, as they are conjugates. Thus, in this case, it is enough to compute one of these stabilizers for a “convenient” $x$.

As the situation of Proposition 15.14 is of particular interest, we make the following definition:

**Definition 15.16.** A set $X$ is said to be a homogeneous space if there is a transitive action $\cdot: G \times X \to X$ of some group $G$ on $X$.

We see that all the spaces of Examples 15.6–15.13, are homogeneous spaces. Another example that will play an important role when we deal with Lie groups is the situation where we have a group $G$, a subgroup $H$ of $G$ (not necessarily normal), and where $X = G/H$, the set of left cosets of $G$ modulo $H$. The group $G$ acts on $G/H$ by left multiplication:

$$a \cdot (gH) = (ag)H,$$
where \( a, g \in G \). This action is clearly transitive and one checks that the stabilizer of \( gH \) is \( gHg^{-1} \). If \( G \) is a topological group and \( H \) is a closed subgroup of \( G \) (see later for an explanation), it turns out that \( G/H \) is Hausdorff. If \( G \) is a Lie group, we obtain a manifold.

Even if \( G \) and \( X \) are topological spaces and the action \( \cdot : G \times X \to X \) is continuous, in general, the space \( G/G_x \) under the quotient topology is \textbf{not} homeomorphic to \( X \).

We will give later sufficient conditions that insure that \( X \) is indeed a topological space or even a manifold. In particular, \( X \) will be a manifold when \( G \) is a Lie group.

In general, an action \( \cdot : G \times X \to X \) is not transitive on \( X \), but for every \( x \in X \), it is transitive on the set
\[
O(x) = G \cdot x = \{ g \cdot x \mid g \in G \}.
\]
Such a set is called the \textit{orbit} of \( x \). The orbits are the equivalence classes of the following equivalence relation:

**Definition 15.17.** Given an action \( \cdot : G \times X \to X \) of some group \( G \) on \( X \), the equivalence relation \( \sim \) on \( X \) is defined so that, for all \( x, y \in X \),
\[
x \sim y \iff y = g \cdot x, \quad \text{for some } g \in G.
\]
For every \( x \in X \), the equivalence class of \( x \) is the \textit{orbit of} \( x \), denoted \( O(x) \) or \( G \cdot x \), with
\[
G \cdot x = O(x) = \{ g \cdot x \mid g \in G \}.
\]
The set of orbits is denoted \( X/G \).

We warn the reader that some authors use the notation \( G \setminus X \) for the the set of orbits \( G \cdot x \), because these orbits can be considered as right orbits, by analogy with right cosets \( Hg \) of a subgroup \( H \) of \( G \).

The orbit space \( X/G \) is obtained from \( X \) by an identification (or merging) process: For every orbit, all points in that orbit are merged into a single point. This akin to the process of forming the identification topology. For example, if \( X = S^2 \) and \( G \) is the group consisting of the restrictions of the two linear maps \( I \) and \( -I \) of \( \mathbb{R}^3 \) to \( S^2 \) (where \( (-I)(x) = -x \) for all \( x \in \mathbb{R}^3 \)), then
\[
X/G = S^2/\{I, -I\} \cong \mathbb{R}P^2.
\]
See Figure 15.4. More generally, if \( S^n \) is the \( n \)-sphere in \( \mathbb{R}^{n+1} \), then we have a bijection between the orbit space \( S^n/\{I, -I\} \) and \( \mathbb{R}P^n \):
\[
S^n/\{I, -I\} \cong \mathbb{R}P^n.
\]
Many manifolds can be obtained in this fashion, including the torus, the Klein bottle, the Möbius band, \textit{etc}.
Since the action of $G$ is transitive on $O(x)$, by Proposition 15.14, we see that for every $x \in X$, we have a bijection

$$O(x) \cong G/G_x.$$ 

As a corollary, if both $X$ and $G$ are finite, for any set $A \subseteq X$ of representatives from every orbit, we have the orbit formula:

$$|X| = \sum_{a \in A} [G: G_a] = \sum_{a \in A} |G|/|G_a|.$$ 

Even if a group action $\cdot : G \times X \to X$ is not transitive, when $X$ is a manifold, we can consider the set of orbits $X/G$, and if the action of $G$ on $X$ satisfies certain conditions, $X/G$ is actually a manifold. Manifolds arising in this fashion are often called orbifolds. In summary, we see that manifolds arise in at least two ways from a group action:

1. As homogeneous spaces $G/G_x$, if the action is transitive.
2. As orbifolds $X/G$ (under certain conditions on the action).

Of course, in both cases, the action must satisfy some additional properties.

For the rest of this section, we reconsider Examples 15.6–15.13 in the context of homogeneous space by determining some stabilizers for those actions.

(a) Consider the action $\cdot : \text{SO}(n) \times S^{n-1} \to S^{n-1}$ of $\text{SO}(n)$ on the sphere $S^{n-1}$ ($n \geq 1$) defined in Example 15.6. Since this action is transitive, we can determine the stabilizer of any convenient element of $S^{n-1}$, say $e_1 = (1,0,\ldots,0)$. In order for any $R \in \text{SO}(n)$ to leave $e_1$ fixed, the first column of $R$ must be $e_1$, so $R$ is an orthogonal matrix of the form

$$R = \begin{pmatrix} 1 & U \\ 0 & S \end{pmatrix}, \quad \text{with} \quad \det(S) = 1,$$

where $U$ is a $1 \times (n-1)$ row vector. As the rows of $R$ must be unit vectors, we see that $U = 0$ and $S \in \text{SO}(n-1)$. Therefore, the stabilizer of $e_1$ is isomorphic to $\text{SO}(n-1)$, and we deduce the bijection

$$\text{SO}(n)/\text{SO}(n-1) \cong S^{n-1}.$$ 

Strictly speaking, $\text{SO}(n-1)$ is not a subgroup of $\text{SO}(n)$, and in all rigor, we should consider the subgroup $\widetilde{\text{SO}}(n-1)$ of $\text{SO}(n)$ consisting of all matrices of the form

$$\begin{pmatrix} 1 & 0 \\ 0 & S \end{pmatrix}, \quad \text{with} \quad \det(S) = 1,$$

and write

$$\text{SO}(n)/\widetilde{\text{SO}}(n-1) \cong S^{n-1}.$$ 

However, it is common practice to identify $\text{SO}(n-1)$ with $\widetilde{\text{SO}}(n-1)$. 
When \( n = 2 \), as \( \text{SO}(1) = \{1\} \), we find that \( \text{SO}(2) \cong S^1 \), a circle, a fact that we already knew. When \( n = 3 \), we find that \( \text{SO}(3)/\text{SO}(2) \cong S^2 \). This says that \( \text{SO}(3) \) is somehow the result of gluing circles to the surface of a sphere (in \( \mathbb{R}^3 \)), in such a way that these circles do not intersect. This is hard to visualize!

A similar argument for the complex unit sphere \( \Sigma^{n-1} \) shows that

\[
\text{SU}(n)/\text{SU}(n-1) \cong \Sigma^{n-1} \cong S^{2n-1}.
\]

Again, we identify \( \text{SU}(n-1) \) with a subgroup of \( \text{SU}(n) \), as in the real case. In particular, when \( n = 2 \), as \( \text{SU}(1) = \{1\} \), we find that

\[
\text{SU}(2) \cong S^3,
\]

that is, the group \( \text{SU}(2) \) is topologically the sphere \( S^3 \)! Actually, this is not surprising if we remember that \( \text{SU}(2) \) is in fact the group of unit quaternions.

(b) We saw in Example 15.7 that the action \( \cdot : \text{SL}(2, \mathbb{R}) \times H \to H \) of the group \( \text{SL}(2, \mathbb{R}) \) on the upper half plane is transitive. Let us find out what the stabilizer of \( z = i \) is. We should have

\[
\frac{ai + b}{ci + d} = i,
\]

that is, \( ai + b = -c + di \), i.e.,

\[(d - a)i = b + c.\]

Since \( a, b, c, d \) are real, we must have \( d = a \) and \( b = -c \). Moreover, \( ad - bc = 1 \), so we get \( a^2 + b^2 = 1 \). We conclude that a matrix in \( \text{SL}(2, \mathbb{R}) \) fixes \( i \) iff it is of the form

\[
\begin{pmatrix}
a & -b \\
\frac{b}{a} & a
\end{pmatrix}, \quad \text{with} \quad a^2 + b^2 = 1.
\]

Clearly, these are the rotation matrices in \( \text{SO}(2) \), and so the stabilizer of \( i \) is \( \text{SO}(2) \). We conclude that

\[
\text{SL}(2, \mathbb{R})/\text{SO}(2) \cong H.
\]

This time we can view \( \text{SL}(2, \mathbb{R}) \) as the result of gluing circles to the upper half plane. This is not so easy to visualize. There is a better way to visualize the topology of \( \text{SL}(2, \mathbb{R}) \) by making it act on the open disk \( D \). We will return to this action in a little while.

(c) Now consider the action of \( \text{SL}(2, \mathbb{C}) \) on \( \mathbb{C} \cup \{\infty\} \cong S^2 \) given in Example 15.8. As it is transitive, let us find the stabilizer of \( z = 0 \). We must have

\[
\frac{b}{d} = 0,
\]

and as \( ad - bc = 1 \), we must have \( b = 0 \) and \( ad = 1 \). Thus the stabilizer of 0 is the subgroup \( \text{SL}(2, \mathbb{C})_0 \) of \( \text{SL}(2, \mathbb{C}) \) consisting of all matrices of the form

\[
\begin{pmatrix}
a & 0 \\
c & a^{-1}
\end{pmatrix}, \quad \text{where} \quad a \in \mathbb{C} - \{0\} \quad \text{and} \quad c \in \mathbb{C}.
\]
We get
\[ \text{SL}(2, \mathbb{C})/\text{SL}(2, \mathbb{C})_0 \cong \mathbb{C} \cup \{\infty\} \cong S^2, \]
but this is not very illuminating.

(d) In Example 15.11 we considered the action \( \cdot : \text{GL}(n) \times \text{SPD}(n) \to \text{SPD}(n) \) of \( \text{GL}(n) \) on \( \text{SPD}(n) \), the set of symmetric positive definite matrices. As this action is transitive, let us find the stabilizer of \( I \). For any \( A \in \text{GL}(n) \), the matrix \( A \) stabilizes \( I \) iff
\[ AIA^\top = AA^\top = I. \]
Therefore the stabilizer of \( I \) is \( \text{O}(n) \), and we find that
\[ \text{GL}(n)/\text{O}(n) = \text{SPD}(n). \]

Observe that if \( \text{GL}^+(n) \) denotes the subgroup of \( \text{GL}(n) \) consisting of all matrices with a strictly positive determinant, then we have an action \( \cdot : \text{GL}^+(n) \times \text{SPD}(n) \to \text{SPD}(n) \) of \( \text{GL}^+(n) \) on \( \text{SPD}(n) \). This action is transitive and we find that the stabilizer of \( I \) is \( \text{SO}(n) \); consequently, we get
\[ \text{GL}^+(n)/\text{SO}(n) = \text{SPD}(n). \]

(e) In Example 15.12 we considered the action \( \cdot : \text{SO}(n+1) \times \mathbb{RP}^n \to \mathbb{RP}^n \) of \( \text{SO}(n+1) \) on the (real) projective space \( \mathbb{RP}^n \). As this action is transitive, let us find the stabilizer of the line \( L = [e_1] \), where \( e_1 = (1, 0, \ldots, 0) \). For any \( R \in \text{SO}(n+1) \), the line \( L \) is fixed iff either \( R(e_1) = e_1 \) or \( R(e_1) = -e_1 \), since \( e_1 \) and \( -e_1 \) define the same line. As \( R \) is orthogonal with \( \det(R) = 1 \), this means that \( R \) is of the form
\[ R = \begin{pmatrix} \alpha & 0 \\ 0 & S \end{pmatrix}, \quad \text{with} \quad \alpha = \pm 1 \quad \text{and} \quad \det(S) = \alpha. \]
But, \( S \) must be orthogonal, so we conclude \( S \in \text{O}(n) \). Therefore, the stabilizer of \( L = [e_1] \) is isomorphic to the group \( \text{O}(n) \), and we find that
\[ \text{SO}(n+1)/\text{O}(n) \cong \mathbb{RP}^n. \]

\[ \text{Strictly speaking,} \, \text{O}(n) \text{ is not a subgroup of} \, \text{SO}(n+1), \text{so the above equation does not make sense. We should write} \]
\[ \text{SO}(n+1)/\tilde{\text{O}}(n) \cong \mathbb{RP}^n, \]
where \( \tilde{\text{O}}(n) \) is the subgroup of \( \text{SO}(n+1) \) consisting of all matrices of the form
\[ \begin{pmatrix} \alpha & 0 \\ 0 & S \end{pmatrix}, \quad \text{with} \quad S \in \text{O}(n), \alpha = \pm 1 \quad \text{and} \quad \det(S) = \alpha. \]
This group is also denoted \( S(\text{O}(1) \times \text{O}(n)) \). However, the common practice is to write \( \text{O}(n) \) instead of \( S(\text{O}(1) \times \text{O}(n)) \).
We should mention that \( \mathbb{RP}^3 \) and \( \text{SO}(3) \) are homeomorphic spaces. This is shown using the quaternions; for example, see Gallier [20], Chapter 8.

A similar argument applies to the action \( \cdot : \text{SU}(n+1) \times \mathbb{CP}^n \to \mathbb{CP}^n \) of \( \text{SU}(n+1) \) on the (complex) projective space \( \mathbb{CP}^n \). We find that

\[
\text{SU}(n+1)/\text{U}(n) \cong \mathbb{CP}^n.
\]

Again, the above is a bit sloppy as \( \text{U}(n) \) is not a subgroup of \( \text{SU}(n+1) \). To be rigorous, we should use the subgroup \( \tilde{\text{U}}(n) \) consisting of all matrices of the form

\[
\begin{pmatrix}
\alpha & 0 \\
0 & S
\end{pmatrix}, \quad \text{with} \ S \in \text{U}(n), \ |\alpha| = 1 \ \text{and} \ \det(S) = \overline{\alpha}.
\]

This group is also denoted \( S(\text{U}(1) \times \text{U}(n)) \). The common practice is to write \( \text{U}(n) \) instead of \( S(\text{U}(1) \times \text{U}(n)) \). In particular, when \( n = 1 \), we find that

\[
\text{SU}(2)/\text{U}(1) \cong \mathbb{CP}^1.
\]

But, we know that \( \text{SU}(2) \cong S^3 \), and clearly \( \text{U}(1) \cong S^1 \). So, again, we find that \( S^3/S^1 \cong \mathbb{CP}^1 \) (we know more, namely, \( S^3/S^1 \cong S^2 \cong \mathbb{CP}^1 \)).

Observe that \( \mathbb{CP}^n \) can also be viewed as the orbit space of the action \( \cdot : S^1 \times S^{2n+1} \to S^{2n+1} \) given by

\[
\lambda \cdot (z_1, \ldots, z_{n+1}) = (\lambda z_1, \ldots, \lambda z_{n+1}),
\]

where \( S^1 = \text{U}(1) \) (the group of complex numbers of modulus 1) and \( S^{2n+1} \) is identified with \( \Sigma^n \).

We now return to Case (b) to give a better picture of \( \text{SL}(2, \mathbb{R}) \). Instead of having \( \text{SL}(2, \mathbb{R}) \) act on the upper half plane, we define an action of \( \text{SL}(2, \mathbb{R}) \) on the open unit disk \( D \) as we did in Example 15.9. Technically, it is easier to consider the group \( \text{SU}(1, 1) \), which is isomorphic to \( \text{SL}(2, \mathbb{R}) \), and to make \( \text{SU}(1, 1) \) act on \( D \). The group \( \text{SU}(1, 1) \) is the group of \( 2 \times 2 \) complex matrices of the form

\[
\begin{pmatrix}
a & b \\
b & \overline{a}
\end{pmatrix}, \quad \text{with} \ a\overline{a} - b\overline{b} = 1.
\]

The reader should check that if we let

\[
g = \begin{pmatrix} 1 & -i \\ 1 & i \end{pmatrix},
\]

then the map from \( \text{SL}(2, \mathbb{R}) \) to \( \text{SU}(1, 1) \) given by

\[
A \mapsto gAg^{-1}
\]
is an isomorphism. Observe that the scaled Möbius transformation associated with $g$ is

$$z \mapsto \frac{z - i}{z + i},$$

which is the holomorphic isomorphism mapping $H$ to $D$ mentioned earlier! We can define a bijection between $\text{SU}(1,1)$ and $S^1 \times D$ given by

$$\begin{pmatrix} a & b \\ b & a \end{pmatrix} \mapsto (a/|a|, b/a).$$

We conclude that $\text{SL}(2, \mathbb{R}) \cong \text{SU}(1,1)$ is topologically an open solid torus (i.e., with the surface of the torus removed). It is possible to further classify the elements of $\text{SL}(2, \mathbb{R})$ into three categories and to have geometric interpretations of these as certain regions of the torus. For details, the reader should consult Carter, Segal and Macdonald [10] or Duistermatt and Kolk [15] (Chapter 1, Section 1.2).

The group $\text{SU}(1,1)$ acts on $D$ by interpreting any matrix in $\text{SU}(1,1)$ as a Möbius transformation; that is,

$$\begin{pmatrix} a & b \\ b & a \end{pmatrix} \mapsto \left( z \mapsto \frac{az + b}{bz + a} \right).$$

The reader should check that these transformations preserve $D$.

Both the upper half-plane and the open disk are models of Lobachevsky’s non-Euclidean geometry (where the parallel postulate fails). They are also models of hyperbolic spaces (Riemannian manifolds with constant negative curvature, see Gallot, Hulin and Lafontaine [21], Chapter III). According to Dubrovin, Fomenko, and Novikov [14] (Chapter 2, Section 13.2), the open disk model is due to Poincaré and the upper half-plane model to Klein, although Poincaré was the first to realize that the upper half-plane is a hyperbolic space.

## 15.4 The Grassmann and Stiefel Manifolds

In this section we introduce two very important homogeneous manifolds, the Grassmann manifolds and the Stiefel manifolds. The Grassmann manifolds are generalizations of projective spaces (real and complex), while the Stiefel manifold are generalizations of $\text{O}(n)$. Both of these manifolds are examples of reductive homogeneous spaces. We begin by defining the Grassmann manifolds $G(k,n)$.

First consider the real case.

**Definition 15.18.** Given any $n \geq 1$, for any $k$ with $0 \leq k \leq n$, the set $G(k,n)$ of all linear $k$-dimensional subspaces of $\mathbb{R}^n$ (also called $k$-planes) is called a **Grassmannian** (or **Grassmann manifold**).
Any \( k \)-dimensional subspace \( U \) of \( \mathbb{R}^n \) is spanned by \( k \) linearly independent vectors \( u_1, \ldots, u_k \) in \( \mathbb{R}^n \); write \( U = \text{span}(u_1, \ldots, u_k) \). We can define an action \( \cdot : \text{O}(n) \times G(k,n) \to G(k,n) \) as follows: For any \( R \in \text{O}(n) \), for any \( U = \text{span}(u_1, \ldots, u_k) \), let 
\[
R \cdot U = \text{span}(Ru_1, \ldots, Ru_k).
\]
We have to check that the above is well defined. If \( U = \text{span}(v_1, \ldots, v_k) \) for any other \( k \) linearly independent vectors \( v_1, \ldots, v_k \), we have 
\[
v_i = \sum_{j=1}^{k} a_{ij} u_j, \quad 1 \leq i \leq k,
\]
for some \( a_{ij} \in \mathbb{R} \), and so 
\[
Rv_i = \sum_{j=1}^{k} a_{ij} Ru_j, \quad 1 \leq i \leq k,
\]
which shows that 
\[
\text{span}(Ru_1, \ldots, Ru_k) = \text{span}(Rv_1, \ldots, Rv_k);
\]
that is, the above action is well defined.

We claim this action is transitive. This is because if \( U \) and \( V \) are any two \( k \)-planes, we may assume that \( U = \text{span}(u_1, \ldots, u_k) \) and \( V = \text{span}(v_1, \ldots, v_k) \), where the \( u_i \)'s form an orthonormal family and similarly for the \( v_i \)'s. Then we can extend these families to orthonormal bases \((u_1, \ldots, u_n)\) and \((v_1, \ldots, v_n)\) on \( \mathbb{R}^n \), and w.r.t. the orthonormal basis \((u_1, \ldots, u_n)\), the matrix of the linear map sending \( u_i \) to \( v_i \) is orthogonal. Hence \( G(k,n) \) is a homogeneous space.

In order to represent \( G(k,n) \) as a quotient space, Proposition 15.14 implies it is enough to find the stabilizer of any \( k \)-plane. Pick \( U = \text{span}(e_1, \ldots, e_k) \), where \((e_1, \ldots, e_n)\) is the canonical basis of \( \mathbb{R}^n \) (i.e., \( e_i = (0, \ldots, 0, 1, 0, \ldots, 0) \), with the 1 in the \( i \)th position). Any \( R \in \text{O}(n) \) stabilizes \( U \) iff \( R \) maps \( e_1, \ldots, e_k \) to \( k \) linearly independent vectors in the subspace \( U = \text{span}(e_1, \ldots, e_k) \), i.e., \( R \) is of the form 
\[
R = \begin{pmatrix} S & 0 \\ 0 & T \end{pmatrix},
\]
where \( S \) is \( k \times k \) and \( T \) is \((n-k) \times (n-k)\). Moreover, as \( R \) is orthogonal, \( S \) and \( T \) must be orthogonal, that is \( S \in \text{O}(k) \) and \( T \in \text{O}(n-k) \). We deduce that the stabilizer of \( U \) is isomorphic to \( \text{O}(k) \times \text{O}(n-k) \) and we find that 
\[
\text{O}(n)/(\text{O}(k) \times \text{O}(n-k)) \cong G(k,n).
\]
It turns out that this makes \( G(k,n) \) into a smooth manifold of dimension 
\[
\frac{n(n-1)}{2} - \frac{k(k-1)}{2} - \frac{(n-k)(n-k-1)}{2} = k(n-k)
\]
called a Grassmannian.

The restriction of the action of $O(n)$ on $G(k, n)$ to $SO(n)$ yields an action
$\cdot: SO(n) \times G(k, n) \rightarrow G(k, n)$ of $SO(n)$ on $G(k, n)$. Then it is easy to see that this action
is transitive and that the stabilizer of the subspace $U$ is isomorphic to the subgroup
$S(O(k) \times O(n - k))$ of $SO(n)$ consisting of the rotations of the form

$$R = \begin{pmatrix} S & 0 \\ 0 & T \end{pmatrix},$$

with $S \in O(k)$, $T \in O(n - k)$ and $\det(S) \det(T) = 1$. Thus, we also have

$$SO(n) / S(O(k) \times O(n - k)) \cong G(k, n).$$

If we recall the projection map of Example 15.12 in Section 15.2, namely $pr: \mathbb{R}^{n+1} - \{0\} \rightarrow \mathbb{RP}^n$, by definition, a $k$-plane in $\mathbb{RP}^n$ is the image under $pr$ of any $(k + 1)$-plane in $\mathbb{R}^{n+1}$. So, for example, a line in $\mathbb{RP}^n$ is the image of a 2-plane in $\mathbb{R}^{n+1}$, and a hyperplane in $\mathbb{RP}^n$ is the image of a hyperplane in $\mathbb{R}^{n+1}$. The advantage of this point of view is that the $k$-planes in $\mathbb{RP}^n$ are arbitrary; that is, they do not have to go through “the origin” (which does not make sense, anyway!). Then we see that we can interpret the Grassmannian, $G(k+1, n+1)$, as a space of “parameters” for the $k$-planes in $\mathbb{RP}^n$. For example, $G(2, n + 1)$ parametrizes the lines in $\mathbb{RP}^n$. In this viewpoint, $G(k+1, n+1)$ is usually denoted $G(k, n)$.

It can be proved (using some exterior algebra) that $G(k, n)$ can be embedded in $\mathbb{RP}^{\binom{n}{k}-1}$. Much more is true. For example, $G(k, n)$ is a projective variety, which means that it can be defined as a subset of $\mathbb{RP}^{\binom{n}{k}-1}$ equal to the zero locus of a set of homogeneous equations. There is even a set of quadratic equations known as the Plücker equations defining $G(k, n)$. In particular, when $n = 4$ and $k = 2$, we have $G(2, 4) \subseteq \mathbb{RP}^5$, and $G(2, 4)$ is defined by a single equation of degree 2. The Grassmannian $G(2, 4) = G(1, 3)$ is known as the Klein quadric. This hypersurface in $\mathbb{RP}^5$ parametrizes the lines in $\mathbb{RP}^3$.

Complex Grassmannians are defined in a similar way, by replacing $\mathbb{R}$ by $\mathbb{C}$ and $O(n)$ by $U(n)$ throughout. The complex Grassmannian $G_C(k, n)$ is a complex manifold as well as a real manifold, and we have

$$U(n) / (U(k) \times U(n-k)) \cong G_C(k, n).$$

As in the case of the real Grassmannians, the action of $U(n)$ on $G_C(k, n)$ yields an action of $SU(n)$ on $G_C(k, n)$, and we get

$$SU(n) / S(U(k) \times U(n-k)) \cong G_C(k, n),$$

where $S(U(k) \times U(n-k))$ is the subgroup of $SU(n)$ consisting of all matrices $R \in SU(n)$ of the form

$$R = \begin{pmatrix} S & 0 \\ 0 & T \end{pmatrix},$$
CHAPTER 15. BASICS OF GROUPS AND GROUP ACTIONS

with \( S \in \mathbf{U}(k), T \in \mathbf{U}(n-k) \) and \( \det(S) \det(T) = 1 \).

Closely related to Grassmannians are the Stiefel manifolds \( S(k, n) \). Again we begin with the real case.

**Definition 15.19.** For any \( n \geq 1 \) and any \( k \) with \( 1 \leq k \leq n \), the set \( S(k, n) \) of all orthonormal \( k \)-frames, that is, of \( k \)-tuples of orthonormal vectors \((u_1, \ldots, u_k)\) with \( u_i \in \mathbb{R}^n \), is called a Stiefel manifold.

Obviously, \( S(1, n) = S^{n-1} \) and \( S(n, n) = \mathbf{O}(n) \), so assume \( k \leq n - 1 \). There is a natural action \( \cdot : \mathbf{SO}(n) \times S(k, n) \to S(k, n) \) of \( \mathbf{SO}(n) \) on \( S(k, n) \) given by

\[
R \cdot (u_1, \ldots, u_k) = (Ru_1, \ldots, Ru_k).
\]

This action is transitive, because if \((u_1, \ldots, u_k)\) and \((v_1, \ldots, v_k)\) are any two orthonormal \( k \)-frames, then they can be extended to orthonormal bases (for example, by Gram-Schmidt) \((u_1, \ldots, u_n)\) and \((v_1, \ldots, v_n)\) with the same orientation (since we can pick \( u_n \) and \( v_n \) so that our bases have the same orientation), and there is a unique orthogonal transformation \( R \in \mathbf{SO}(n) \) such that \( Ru_i = v_i \) for \( i = 1, \ldots, n \).

In order to apply Proposition 15.14, we need to find the stabilizer of the orthonormal \( k \)-frame \((e_1, \ldots, e_k)\) consisting of the first canonical basis vectors of \( \mathbb{R}^n \). A matrix \( R \in \mathbf{SO}(n) \) stabilizes \((e_1, \ldots, e_k)\) iff it is of the form

\[
R = \begin{pmatrix} I_k & 0 \\ 0 & S \end{pmatrix}
\]

where \( S \in \mathbf{SO}(n-k) \). Therefore, for \( 1 \leq k \leq n-1 \), we have

\[
\mathbf{SO}(n)/\mathbf{SO}(n-k) \cong S(k, n).
\]

This makes \( S(k, n) \) a smooth manifold of dimension

\[
\frac{n(n-1)}{2} - \frac{(n-k)(n-k-1)}{2} = nk - k(k+1) = k(n-k) + \frac{k(k-1)}{2}.
\]

**Remark:** It should be noted that we can define another type of Stiefel manifolds, denoted by \( V(k, n) \), using linearly independent \( k \)-tuples \((u_1, \ldots, u_k)\) that do not necessarily form an orthonormal system. In this case, there is an action \( \cdot : \mathbf{GL}(n, \mathbb{R}) \times V(k, n) \to V(k, n) \), and the stabilizer \( H \) of the first \( k \) canonical basis vectors \((e_1, \ldots, e_k)\) is a closed subgroup of \( \mathbf{GL}(n, \mathbb{R}) \), but it doesn’t have a simple description (see Warner [43], Chapter 3). We get an isomorphism

\[
V(k, n) \cong \mathbf{GL}(n, \mathbb{R})/H.
\]

The version of the Stiefel manifold \( S(k, n) \) using orthonormal frames is sometimes denoted by \( V^0(k, n) \) (Milnor and Stasheff [30] use the notation \( V^0_k(\mathbb{R}^n) \)). Beware that the notation is not standardized. Certain authors use \( V(k, n) \) for what we denote by \( S(k, n) \)!
Complex Stiefel manifolds are defined in a similar way by replacing \( \mathbb{R} \) by \( \mathbb{C} \) and \( \text{SO}(n) \) by \( \text{SU}(n) \). For \( 1 \leq k \leq n - 1 \), the complex Stiefel manifold \( S_{\mathbb{C}}(k, n) \) is isomorphic to the quotient
\[
\text{SU}(n)/\text{SU}(n-k) \cong S_{\mathbb{C}}(k, n).
\]
If \( k = 1 \), we have \( S_{\mathbb{C}}(1, n) = S^{2n-1} \), and if \( k = n \), we have \( S_{\mathbb{C}}(n, n) = \text{U}(n) \).

The Grassmannians can also be viewed as quotient spaces of the Stiefel manifolds. Every orthonomal \( k \)-frame \( (u_1, \ldots, u_k) \) can be represented by an \( n \times k \) matrix \( Y \) over the canonical basis of \( \mathbb{R}^n \), and such a matrix \( Y \) satisfies the equation
\[
Y^\top Y = I.
\]

We have a right action \( \cdot : S(k, n) \times \text{O}(k) \rightarrow S(k, n) \) given by
\[
Y \cdot R = YR,
\]
for any \( R \in \text{O}(k) \). This action is well defined since
\[
(YR)^\top YR = R^\top Y^\top YR = I.
\]
However, this action is not transitive (unless \( k = 1 \)), but the orbit space \( S(k, n)/\text{O}(k) \) is isomorphic to the Grassmannian \( G(k, n) \), so we can write
\[
G(k, n) \cong S(k, n)/\text{O}(k).
\]
Similarly, the complex Grassmannian is isomorphic to the orbit space \( S_{\mathbb{C}}(k, n)/\text{U}(k) \):
\[
G_{\mathbb{C}}(k, n) \cong S_{\mathbb{C}}(k, n)/\text{U}(k).
\]
Bibliography


303


