Mathematical Analysis
Volume II

Teo Lee Peng

Mathematical Analysis
Volume II

Teo Lee Peng

July 3, 2025

Contents

Contents i

Preface iv

Chapter 1 Euclidean Spaces 1
1.1 The Euclidean Space $\mathbb{R}^n$ as a Vector Space 1
1.2 Convergence of Sequences in $\mathbb{R}^n$ 23
1.3 Open Sets and Closed Sets 33
1.4 Interior, Exterior, Boundary and Closure 46
1.5 Limit Points and Isolated Points 59

Chapter 2 Limits of Multivariable Functions and Continuity 66
2.1 Multivariable Functions 66
2.1.1 Polynomials and Rational Functions 66
2.1.2 Component Functions of a Mapping 68
2.1.3 Invertible Mappings 69
2.1.4 Linear Transformations 70
2.1.5 Quadratic Forms 74
2.2 Limits of Functions 79
2.3 Continuity 92
2.4 Uniform Continuity 121
2.5 Contraction Mapping Theorem 127

Chapter 3 Continuous Functions on Connected Sets and Compact Sets 132
3.1 Path-Connectedness and Intermediate Value Theorem 132
3.2 Connectedness and Intermediate Value Property 147
3.3 Sequential Compactness and Compactness 161
3.4 Applications of Compactness 181
3.4.1 The Extreme Value Theorem 181
3.4.2 Distance Between Sets 184

3.4.3 Uniform Continuity . . . 191
3.4.4 Linear Transformations and Quadratic Forms . . . 192
3.4.5 Lebesgue Number Lemma . . . 195

Chapter 4 Differentiating Functions of Several Variables 201
4.1 Partial Derivatives . . . 201
4.2 Differentiability and First Order Approximation . . . 221
4.2.1 Differentiability . . . 221
4.2.2 First Order Approximations . . . 233
4.2.3 Tangent Planes . . . 237
4.2.4 Directional Derivatives . . . 238
4.3 The Chain Rule and the Mean Value Theorem . . . 248
4.4 Second Order Approximations . . . 263
4.5 Local Extrema . . . 271

Chapter 5 The Inverse and Implicit Function Theorems 285
5.1 The Inverse Function Theorem . . . 285
5.2 The Proof of the Inverse Function Theorem . . . 298
5.3 The Implicit Function Theorem . . . 309
5.4 Extrema Problems and the Method of Lagrange Multipliers . . . 329

Chapter 6 Multiple Integrals 343
6.1 Riemann Integrals . . . 344
6.2 Properties of Riemann Integrals . . . 376
6.3 Jordan Measurable Sets and Riemann Integrable Functions . . . 389
6.4 Iterated Integrals and Fubini's Theorem . . . 431
6.5 Change of Variables Theorem . . . 450
6.5.1 Translations and Linear Transformations . . . 454
6.5.2 Polar Coordinates . . . 466
6.5.3 Spherical Coordinates . . . 477
6.5.4 Other Examples . . . 482
6.6 Proof of the Change of Variables Theorem . . . 487
6.7 Some Important Integrals and Their Applications . . . 509

Chapter 7 Fourier Series and Fourier Transforms **517**
7.1 Orthogonal Systems of Functions and Fourier Series 518
7.2 The Pointwise Convergence of a Fourier Series 540
7.3 The L^2 Convergence of a Fourier Series 556
7.4 The Uniform Convergence of a Trigonometric Series 570
7.5 Fourier Transforms 586

Appendix A Sylvester's Criterion **615**

Appendix B Volumes of Parallelepipeds **622**

Appendix C Riemann Integrability **629**

References **642**

Preface

Mathematical analysis is a standard course which introduces students to rigorous reasonings in mathematics, as well as the theories needed for advanced analysis courses. It is a compulsory course for all mathematics majors. It is also strongly recommended for students that major in computer science, physics, data science, financial analysis, and other areas that require a lot of analytical skills. Some standard textbooks in mathematical analysis include the classical one by Apostol [Apo74] and Rudin [Rud76], and the modern one by Bartle [BS92], Fitzpatrick [Fit09], Abbott [Abb15], Tao [Tao16, Tao14] and Zorich [Zor15, Zor16].

This book is the second volume of the textbooks intended for a one-year course in mathematical analysis. We introduce the fundamental concepts in a pedagogical way. Lots of examples are given to illustrate the theories. We assume that students are familiar with the material of calculus such as those in the book [SCW20]. Thus, we do not emphasize on the computation techniques. Emphasis is put on building up analytical skills through rigorous reasonings.

Besides calculus, it is also assumed that students have taken introductory courses in discrete mathematics and linear algebra, which covers topics such as logic, sets, functions, vector spaces, inner products, and quadratic forms. Whenever needed, these concepts would be briefly revised.

In this book, we have defined all the mathematical terms we use carefully. While most of the terms have standard definitions, some of the terms may have definitions defer from authors to authors. The readers are advised to check the definitions of the terms used in this book when they encounter them. This can be easily done by using the search function provided by any PDF viewer. The readers are also encouraged to fully utilize the hyper-referencing provided.

Teo Lee Peng

Chapter 1

Euclidean Spaces

In this second volume of mathematical analysis, we study functions defined on subsets of $\mathbb{R}^n$. For this, we need to study the structure and topology of $\mathbb{R}^n$ first. We start by a revision on $\mathbb{R}^n$ as a vector space.

In the sequel, n is a fixed positive integer reserved to be used for $\mathbb{R}^n$.

1.1 The Euclidean Space $\mathbb{R}^n$ as a Vector Space

If S_1, S_2, ..., S_n are sets, the cartesian product of these n sets is defined as the set

$$S = S_1 \times \cdots \times S_n = \prod_{i=1}^{n} S_i = \{(a_1, \ldots, a_n) \,|\, a_i \in S_i, 1 \leq i \leq n\}$$

that contains all n-tuples $(a_1, \ldots, a_n)$, where $a_i \in S_i$ for all $1 \leq i \leq n$.

The set $\mathbb{R}^n$ is the cartesian product of n copies of $\mathbb{R}$. Namely,

$$\mathbb{R}^n = \{(x_1, x_2, \ldots, x_n) \,|\, x_1, x_2, \ldots, x_n \in \mathbb{R}\}\,.$$

The point $(x_1, x_2, \ldots, x_n)$ is denoted as $\mathbf{x}$, whereas $x_1, x_2, \ldots, x_n$ are called the components of the point $\mathbf{x}$. We can define an addition and a scalar multiplication on $\mathbb{R}^n$. If $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ are in $\mathbb{R}^n$, the addition of $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\mathbf{x} + \mathbf{y} = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n).$$

In other words, it is a componentwise addition. Given a real number α, the scalar multiplication of α with $\mathbf{x}$ is given by the componentwise multiplication

$$\alpha\mathbf{x} = (\alpha x_1, \alpha x_2, \ldots, \alpha x_n).$$

The set $\mathbb{R}^n$ with the addition and scalar multiplication operations is a vector space. It satisfies the 10 axioms for a real vector space V.

The 10 Axioms for a Real Vector Space V

Let V be a set that is equipped with two operations – the addition and the scalar multiplication. For any two vectors $\mathbf{u}$ and $\mathbf{v}$ in V, their addition is denoted by $\mathbf{u}+\mathbf{v}$. For a vector $\mathbf{u}$ in V and a scalar $\alpha \in \mathbb{R}$, the scalar multiplication of $\mathbf{v}$ by α is denoted by $\alpha\mathbf{v}$. We say that V with the addition and scalar multiplication is a real vector space provided that the following 10 axioms are satisfied for any $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ in V, and any α and β in $\mathbb{R}$.

Axiom 1 If $\mathbf{u}$ and $\mathbf{v}$ are in V, then $\mathbf{u}+\mathbf{v}$ is in V.

Axiom 2 $\mathbf{u}+\mathbf{v}=\mathbf{v}+\mathbf{u}$.

Axiom 3 $(\mathbf{u}+\mathbf{v})+\mathbf{w}=\mathbf{u}+(\mathbf{v}+\mathbf{w})$.

Axiom 4 There is a zero vector $\mathbf{0}$ in V such that

$$\mathbf{0}+\mathbf{v}=\mathbf{v}=\mathbf{v}+\mathbf{0} \qquad \text{for all } \mathbf{v}\in V.$$

Axiom 5 For any $\mathbf{v}$ in V, there is a vector $\mathbf{w}$ in V such that

$$\mathbf{v}+\mathbf{w}=\mathbf{0}=\mathbf{w}+\mathbf{v}.$$

The vector $\mathbf{w}$ satisfying this equation is called the *negative* of $\mathbf{v}$, and is denoted by $-\mathbf{v}$.

Axiom 6 For any $\mathbf{v}$ in V, and any $\alpha\in\mathbb{R}$, $\alpha\mathbf{v}$ is in V.

Axiom 7 $\alpha(\mathbf{u}+\mathbf{v})=\alpha\mathbf{u}+\alpha\mathbf{v}$.

Axiom 8 $(\alpha+\beta)\mathbf{v}=\alpha\mathbf{v}+\beta\mathbf{v}$.

Axiom 9 $\alpha(\beta\mathbf{v})=(\alpha\beta)\mathbf{v}$.

Axiom 10 $1\mathbf{v}=\mathbf{v}$.

$\mathbb{R}^n$ is a real vector space. The zero vector is the point $\mathbf{0}=(0,0,\ldots,0)$ with all components equal to 0. Sometimes we also call a point $\mathbf{x}=(x_1,\ldots,x_n)$ in

$\mathbb{R}^n$ a vector, and identify it as the vector from the origin $\mathbf{0}$ to the point $\mathbf{x}$.

Definition 1.1 Standard Unit Vectors

In $\mathbb{R}^n$, there are n standard unit vectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$ given by

$$\mathbf{e}_1 = (1, 0, \ldots, 0),\ \mathbf{e}_2 = (0, 1, \ldots, 0),\ \cdots,\ \mathbf{e}_n = (0, \ldots, 0, 1).$$

Let us review some concepts from linear algebra which will be useful later. Given that $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are vectors in a vector space V, a linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_k$ is a vector $\mathbf{v}$ in V of the form

$$\mathbf{v} = c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k$$

for some scalars $c_1, \ldots, c_k$, which are known as the coefficients of the linear combination.

A subspace of a vector space V is a subset of V that is itself a vector space. There is a simple way to construct subspaces.

Proposition 1.1

Let V be a vector space, and let $\mathbf{v}_1, \ldots, \mathbf{v}_k$ be vectors in V. The subset

$$W = \{c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k \,|\, c_1, \ldots, c_k \in \mathbb{R}\}$$

of V that contains all linear combinations of $\mathbf{v}_1, \ldots, \mathbf{v}_k$ is itself a vector space. It is called the subspace of V *spanned* by $\mathbf{v}_1, \ldots, \mathbf{v}_k$.

Example 1.1

In $\mathbb{R}^3$, the subspace spanned by the vectors $\mathbf{e}_1 = (1, 0, 0)$ and $\mathbf{e}_3 = (0, 0, 1)$ is the set W that contains all points of the form

$$x(1, 0, 0) + z(0, 0, 1) = (x, 0, z),$$

which is the xz-plane.

Next, we recall the concept of linear independence.

Definition 1.2 Linear Independence

Let V be a vector space, and let $\mathbf{v}_1, \ldots, \mathbf{v}_k$ be vectors in V. We say that the set $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is a linearly independent set of vectors, or the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are linearly independent, if the only k-tuple of real numbers $(c_1, \ldots, c_k)$ which satisfies

$$c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k = \mathbf{0}$$

is the *trivial* k-tuple $(c_1, \ldots, c_k) = (0, \ldots, 0)$.

Example 1.2

In $\mathbb{R}^n$, the standard unit vectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$ are linearly independent.

Example 1.3

If V is a vector space, a vector $\mathbf{v}$ in V is linearly independent if and only if $\mathbf{v} \neq \mathbf{0}$.

Example 1.4

Let V be a vector space. Two vectors $\mathbf{u}$ and $\mathbf{v}$ in V are linearly independent if and only if $\mathbf{u} \neq \mathbf{0}$, $\mathbf{v} \neq \mathbf{0}$, and there does not exists a constant α such that $\mathbf{v} = \alpha\mathbf{u}$.

Let us recall the following definition for two vectors to be parallel.

Definition 1.3 Parallel Vectors

Let V be a vector space. Two vectors $\mathbf{u}$ and $\mathbf{v}$ in V are parallel if either $\mathbf{u} = \mathbf{0}$ or there exists a constant α such that $\mathbf{v} = \alpha\mathbf{u}$.

In other words, two vectors $\mathbf{u}$ and $\mathbf{v}$ in V are linearly independent if and only if they are not parallel.

Example 1.5

If $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is a linearly independent set of vectors, then for any $S' \subset S$, S' is also a linearly independent set of vectors.

Now we discuss the concept of dimension and basis.

Definition 1.4 Dimension and Basis

Let V be a vector space, and let W be a subspace of V. If W can be spanned by k linearly independent vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ in V, we say that W has dimension k. The set $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is called a basis of W.

Example 1.6

In $\mathbb{R}^n$, the n standard unit vectors $\mathbf{e}_1$, ..., $\mathbf{e}_n$ are linearly independent and they span $\mathbb{R}^n$. Hence, the dimension of $\mathbb{R}^n$ is n.

Example 1.7

In $\mathbb{R}^3$, the subspace spanned by the two linearly independent vectors $\mathbf{e}_1 = (1, 0, 0)$ and $\mathbf{e}_3 = (0, 0, 1)$ has dimension 2.

Next, we introduce the translate of a set.

Definition 1.5 Translate of a Set

If A is a subset of $\mathbb{R}^n$, $\mathbf{u}$ is a point in $\mathbb{R}^n$, the translate of the set A by the vector $\mathbf{u}$ is the set

$$A + \mathbf{u} = \{\mathbf{a} + \mathbf{u} \,|\, \mathbf{a} \in A\}.$$

Example 1.8

In $\mathbb{R}^3$, the translate of the set $A = \{(x, y, 0) \,|\, x, y \in \mathbb{R}\}$ by the vector $\mathbf{u} = (0, 0, -2)$ is the set

$$B = A + \mathbf{u} = \{(x, y, -2) \,|\, x, y \in \mathbb{R}\}.$$

In $\mathbb{R}^n$, the lines and the planes are of particular interest. They are closely

related to the concept of subspaces.

Definition 1.6 Lines in $\mathbb{R}^n$

A line L in $\mathbb{R}^n$ is a translate of a subspace of $\mathbb{R}^n$ that has dimension 1. As a set, it contains all the points $\mathbf{x}$ of the form

$$\mathbf{x} = \mathbf{x}_0 + t\mathbf{v}, \qquad t \in \mathbb{R},$$

where $\mathbf{x}_0$ is a fixed point in $\mathbb{R}^n$, and $\mathbf{v}$ is a nonzero vector in $\mathbb{R}^n$. The equation $\mathbf{x} = \mathbf{x}_0 + t\mathbf{v}$, $t \in \mathbb{R}$, is known as the parametric equation of the line.

A line is determined by two points.

Example 1.9

Given two distinct points $\mathbf{x}_1$ and $\mathbf{x}_2$ in $\mathbb{R}^n$, the line L that passes through these two points have parametric equation given by

$$\mathbf{x} = \mathbf{x}_1 + t(\mathbf{x}_2 - \mathbf{x}_1), \qquad t \in \mathbb{R}.$$

When $0 \leq t \leq 1$, $\mathbf{x} = \mathbf{x}_1 + t(\mathbf{x}_2 - \mathbf{x}_1)$ describes all the points on the line segment with $\mathbf{x}_1$ and $\mathbf{x}_2$ as endpoints.

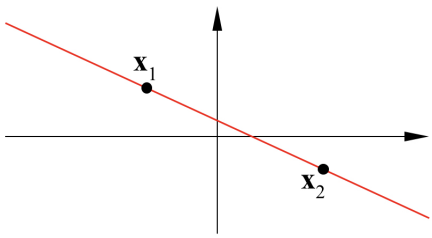


Figure 1.1: A Line between two points.

Definition 1.7 Planes in $\mathbb{R}^n$

A plane W in $\mathbb{R}^n$ is a translate of a subspace of dimension 2. As a set, it contains all the points $\mathbf{x}$ of the form

$$\mathbf{x} = \mathbf{x}_0 + t_1\mathbf{v}_1 + t_2\mathbf{v}_2, \qquad t_1, t_2 \in \mathbb{R},$$

where $\mathbf{x}_0$ is a fixed point in $\mathbb{R}^n$, and $\mathbf{v}_1$ and $\mathbf{v}_2$ are two linearly independent vectors in $\mathbb{R}^n$.

Besides being a real vector space, $\mathbb{R}^n$ has an additional structure. Its definition is motivated as follows. Let $P(x_1, x_2, x_3)$ and $Q(y_1, y_2, y_3)$ be two points in $\mathbb{R}^3$. By Pythagoras theorem, the distance between P and Q is given by

$$PQ = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}.$$

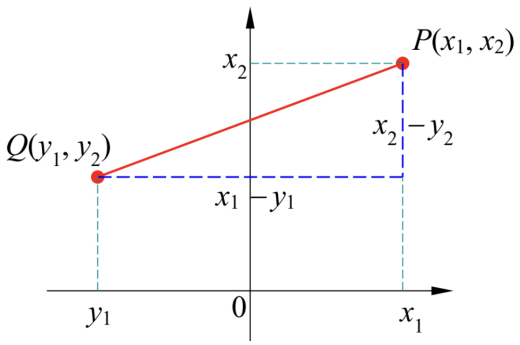


Figure 1.2: Distance between two points in $\mathbb{R}^2$.

Consider the triangle OPQ with vertices O, P, Q, where O is the origin. Then

$$OP = \sqrt{x_1^2 + x_2^2 + x_3^2}, \qquad OQ = \sqrt{y_1^2 + y_2^2 + y_3^2}.$$

Let θ be the minor angle between OP and OQ. By cosine rule,

$$PQ^2 = OP^2 + OQ^2 - 2 \times OP \times OQ \times \cos\theta.$$

A straightforward computation gives

$$OP^2 + OQ^2 - PQ^2 = 2(x_1y_1 + x_2y_2 + x_3y_3).$$

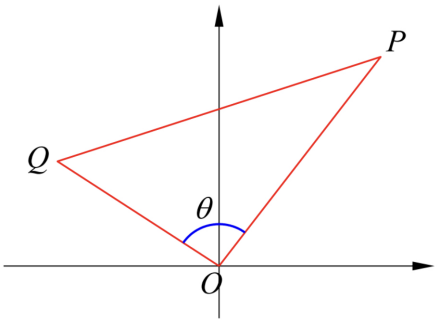


Figure 1.3: Cosine rule.

Hence,

$$\cos\theta = \frac{x_1y_1 + x_2y_2 + x_3y_3}{\sqrt{x_1^2 + x_2^2 + x_3^2}\,\sqrt{y_1^2 + y_2^2 + y_3^2}}. \tag{1.1}$$

It is a quotient of $x_1y_1 + x_2y_2 + x_3y_3$ by the product of the lengths of OP and OQ.

Generalizing the expression $x_1y_1 + x_2y_2 + x_3y_3$ from $\mathbb{R}^3$ to $\mathbb{R}^n$ defines the dot product. For any two vectors $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ in $\mathbb{R}^n$, the dot product of $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{n} x_i y_i = x_1y_1 + x_2y_2 + \cdots + x_ny_n.$$

This is a special case of an inner product.

Definition 1.8 Inner Product Space

A real vector space V is an inner product space if for any two vectors $\mathbf{u}$ and $\mathbf{v}$ in V, an inner product $\langle \mathbf{u}, \mathbf{v} \rangle$ of $\mathbf{u}$ and $\mathbf{v}$ is defined, and the following conditions for any $\mathbf{u}, \mathbf{v}, \mathbf{w}$ in V and $\alpha, \beta \in \mathbb{R}$ are satisfied.

1. $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$.
2. $\langle \alpha\mathbf{u} + \beta\mathbf{v}, \mathbf{w} \rangle = \alpha\langle \mathbf{u}, \mathbf{w} \rangle + \beta\langle \mathbf{v}, \mathbf{w} \rangle$.
3. $\langle \mathbf{v}, \mathbf{v} \rangle \geq 0$ and $\langle \mathbf{v}, \mathbf{v} \rangle = 0$ if and only if $\mathbf{v} = \mathbf{0}$.

Proposition 1.2 Euclidean Inner Product on $\mathbb{R}^n$

On $\mathbb{R}^n$,

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n.$$

defines an inner product, called the standard inner product or the Euclidean inner product.

Definition 1.9 Euclidean Space

The vector space $\mathbb{R}^n$ with the Euclidean inner product is called the Euclidean n-space.

In the future, when we do not specify, $\mathbb{R}^n$ always means the Euclidean n-space.

One can deduce some useful identities from the three axioms of an inner product space.

Proposition 1.3

If V is an inner product space, then the following holds.

(a) For any $\mathbf{v} \in V$, $\langle \mathbf{0}, \mathbf{v} \rangle = 0 = \langle \mathbf{v}, \mathbf{0} \rangle$.

(b) For any vectors $\mathbf{v}_1, \cdots, \mathbf{v}_k, \mathbf{w}_1, \cdots, \mathbf{w}_l$ in V, and for any real numbers $\alpha_1, \cdots, \alpha_k, \beta_1, \cdots, \beta_l$,

$$\left\langle \sum_{i=1}^{k} \alpha_i \mathbf{v}_i, \sum_{j=1}^{l} \beta_j \mathbf{w}_j \right\rangle = \sum_{i=1}^{k} \sum_{j=1}^{l} \alpha_i \beta_j \langle \mathbf{v}_i, \mathbf{w}_j \rangle.$$

Given that V is an inner product space, $\langle \mathbf{v}, \mathbf{v} \rangle \geq 0$ for any $\mathbf{v}$ in V. For example, for any $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ in $\mathbb{R}^n$, under the Euclidean inner product,

$$\langle \mathbf{x}, \mathbf{x} \rangle = \sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + \cdots + x_n^2 \geq 0.$$

When $n = 3$, the length of the vector OP from the point $O(0, 0, 0)$ to the point

$P(x_1, x_2, x_3)$ is

$$OP = \sqrt{x_1^2 + x_2^2 + x_3^2} = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}, \qquad \text{where } \mathbf{x} = (x_1, x_2, x_3).$$

This motivates us to define to norm of a vector in an inner product space as follows.

Definition 1.10 Norm of a Vector

Given that V is an inner product space, the norm of a vector $\mathbf{v}$ is defined as

$$\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}.$$

The norm of a vector in an inner product space satisfies some properties, which follow from the axioms for an inner product space.

Proposition 1.4

Let V be an inner product space.

1. For any $\mathbf{v}$ in V, $\|\mathbf{v}\| \geq 0$ and $\|\mathbf{v}\| = 0$ if and only if $\mathbf{v} = \mathbf{0}$.
2. For any $\alpha \in \mathbb{R}$ and $\mathbf{v} \in V$, $\|\alpha \mathbf{v}\| = |\alpha| \, \|\mathbf{v}\|$.

Motivated by the distance between two points in $\mathbb{R}^3$, we make the following definition.

Definition 1.11 Distance Between Two Points

Given that V is an inner product space, the distance between $\mathbf{u}$ and $\mathbf{v}$ in V is defined as

$$d(\mathbf{u}, \mathbf{v}) = \|\mathbf{v} - \mathbf{u}\| = \sqrt{\langle \mathbf{v} - \mathbf{u}, \mathbf{v} - \mathbf{u} \rangle}.$$

For example, the distance between the points $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ in the Euclidean space $\mathbb{R}^n$ is

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}.$$

For analysis in $\mathbb{R}$, an important inequality is the triangle inequality which says that $|x+y| \leq |x| + |y|$ for any x and y in $\mathbb{R}$. To generalize this inequality to $\mathbb{R}^n$, we need the celebrated Cauchy-Schwarz inequality. It holds on any inner product space.

Proposition 1.5 Cauchy-Schwarz Inequality

Given that V is an inner product space, for any $\mathbf{u}$ and $\mathbf{v}$ in V,

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \ \leq \ \|\mathbf{u}\| \, \|\mathbf{v}\|.$$

The equality holds if and only if $\mathbf{u}$ and $\mathbf{v}$ are parallel.

Proof

It is obvious that if either $\mathbf{u} = \mathbf{0}$ or $\mathbf{v} = \mathbf{0}$,

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \ = \ 0 \ = \ \|\mathbf{u}\| \, \|\mathbf{v}\|,$$

and so the equality holds.
Now assume that both $\mathbf{u}$ and $\mathbf{v}$ are nonzero vectors. Consider the quadratic function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(t) = \|t\mathbf{u} - \mathbf{v}\|^2 = \langle t\mathbf{u} - \mathbf{v}, t\mathbf{u} - \mathbf{v} \rangle.$$

Notice that $f(t) = at^2 + bt + c$, where

$$a = \langle \mathbf{u}, \mathbf{u} \rangle = \|\mathbf{u}\|^2, \quad b = -2\langle \mathbf{u}, \mathbf{v} \rangle, \quad c = \langle \mathbf{v}, \mathbf{v} \rangle = \|\mathbf{v}\|^2.$$

The 3rd axiom of an inner product says that $f(t) \geq 0$ for all $t \in \mathbb{R}$. Hence, we must have $b^2 - 4ac \leq 0$. This gives

$$\langle \mathbf{u}, \mathbf{v} \rangle^2 \leq \|\mathbf{u}\|^2 \|\mathbf{v}\|^2.$$

Thus, we obtain the Cauchy-Schwarz inequality

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \ \leq \ \|\mathbf{u}\| \, \|\mathbf{v}\|.$$

The equality holds if and only if $b^2 - 4ac = 0$. The latter means that $f(t) = 0$ for some $t = \alpha$, which can happen if and only if

$$\alpha\mathbf{u} - \mathbf{v} = \mathbf{0},$$

or equivalently, $\mathbf{v} = \alpha\mathbf{u}$.

Now we can prove the triangle inequality.

Proposition 1.6 Triangle Inequality

Let V be an inner product space. For any vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ in V,

$$\|\mathbf{v}_1 + \mathbf{v}_2 + \cdots + \mathbf{v}_k\| \leq \|\mathbf{v}_1\| + \|\mathbf{v}_2\| + \cdots + \|\mathbf{v}_k\|.$$

Proof

It is sufficient to prove the statement when $k = 2$. The general case follows from induction. Given $\mathbf{v}_1$ and $\mathbf{v}_2$ in V,

$$\begin{aligned}
\|\mathbf{v}_1 + \mathbf{v}_2\|^2 &= \langle \mathbf{v}_1 + \mathbf{v}_2, \mathbf{v}_1 + \mathbf{v}_2 \rangle \\
&= \langle \mathbf{v}_1, \mathbf{v}_1 \rangle + 2\langle \mathbf{v}_1, \mathbf{v}_2 \rangle + \langle \mathbf{v}_2, \mathbf{v}_2 \rangle \\
&\leq \|\mathbf{v}_1\|^2 + 2\|\mathbf{v}_1\|\|\mathbf{v}_2\| + \|\mathbf{v}_2\|^2 \\
&= \left(\|\mathbf{v}_1\| + \|\mathbf{v}_2\|\right)^2.
\end{aligned}$$

This proves that

$$\|\mathbf{v}_1 + \mathbf{v}_2\| \leq \|\mathbf{v}_1\| + \|\mathbf{v}_2\|.$$

From the triangle inequality, we can deduce the following.

Corollary 1.7

Let V be an inner product space. For any vectors $\mathbf{u}$ and $\mathbf{v}$ in V,

$$\big|\|\mathbf{u}\| - \|\mathbf{v}\|\big| \leq \|\mathbf{u} - \mathbf{v}\|.$$

Express in terms of distance, the triangle inequality takes the following form.

Proposition 1.8 Triangle Inequality

Let V be an inner product space. For any three points $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ in V,

$$d(\mathbf{v}_1, \mathbf{v}_2) \leq d(\mathbf{v}_1, \mathbf{v}_3) + d(\mathbf{v}_2, \mathbf{v}_3).$$

More generally, if $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ are k vectors in V, then

$$d(\mathbf{v}_1, \mathbf{v}_k) \leq \sum_{i=2}^{k} d(\mathbf{v}_{i-1}, \mathbf{v}_i) = d(\mathbf{v}_1, \mathbf{v}_2) + \cdots + d(\mathbf{v}_{k-1}, \mathbf{v}_k).$$

Since we can define the distance function on an inner product space, inner product space is a special case of metric spaces.

Definition 1.12 Metric Space

Let X be a set, and let $d : X \times X \to \mathbb{R}$ be a function defined on $X \times X$. We say that d is a metric on X provided that the following conditions are satisfied.

1. For any x and y in X, $d(x, y) \geq 0$, and $d(x, y) = 0$ if and only if $x = y$.
2. $d(x, y) = d(y, x)$ for any x and y in X.
3. For any x, y and z in X, $d(x, y) \leq d(x, z) + d(y, z)$.

If d is a metric on X, we say that (X, d) is a metric space.

Metric spaces play important roles in advanced analysis. If V is an innner product space, it is a metric space with metric

$$d(\mathbf{u}, \mathbf{v}) = \|\mathbf{v} - \mathbf{u}\|.$$

Using the Cauchy-Schwarz inequality, one can generalize the concept of angles to any two vectors in a real inner product space. If $\mathbf{u}$ and $\mathbf{v}$ are two nonzero vectors in a real inner product space V, Cauchy-Schwarz inequality implies that

$$\frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\| \, \|\mathbf{v}\|}$$

is a real number between -1 and 1. Generalizing the formula (1.1), we define the angle θ between $\mathbf{u}$ and $\mathbf{v}$ as

$$\theta = \cos^{-1} \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\| \, \|\mathbf{v}\|}.$$

This is an angle between 0° and 180°. A necessary and sufficient condition for two vectors $\mathbf{u}$ and $\mathbf{v}$ to make a 90° angle is $\langle \mathbf{u}, \mathbf{v} \rangle = 0$.

Definition 1.13 Orthogonality

Let V be a real inner product space. We say that the two vectors $\mathbf{u}$ and $\mathbf{v}$ in V are orthogonal if $\langle \mathbf{u}, \mathbf{v} \rangle = 0$.

Lemma 1.9 Generalized Pythagoras Theorem

Let V be an inner product space. If $\mathbf{u}$ and $\mathbf{v}$ are orthogonal vectors in V, then

$$\|\mathbf{u} + \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2.$$

Now we discuss the projection theorem.

Theorem 1.10 Projection Theorem

Let V be an inner product space, and let $\mathbf{w}$ be a nonzero vector in V. If $\mathbf{v}$ is a vector in V, there is a unique way to write $\mathbf{v}$ as a sum of two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, such that $\mathbf{v}_1$ is parallel to $\mathbf{w}$ and $\mathbf{v}_2$ is orthogonal to $\mathbf{w}$. Moreover, for any real number α,

$$\|\mathbf{v} - \alpha \mathbf{w}\| \geq \|\mathbf{v} - \mathbf{v_1}\|,$$

and the equality holds if and only if α is equal to the unique real number β such that $\mathbf{v}_1 = \beta \mathbf{w}$.

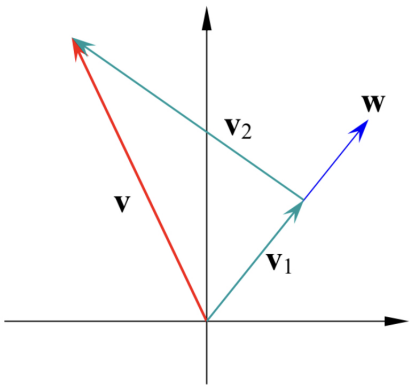


Figure 1.4: The projection theorem.

Proof

Assume that $\mathbf{v}$ can be written as a sum of two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, such that $\mathbf{v}_1$ is parallel to $\mathbf{w}$ and $\mathbf{v}_2$ is orthogonal to $\mathbf{w}$. Since $\mathbf{w}$ is nonzero, there is a real number β such that $\mathbf{v}_1 = \beta\mathbf{w}$. Since $\mathbf{v}_2 = \mathbf{v} - \mathbf{v}_1 = \mathbf{v} - \beta\mathbf{w}$ is orthogonal to $\mathbf{w}$, we have

$$0 = \langle \mathbf{v} - \beta\mathbf{w}, \mathbf{w} \rangle = \langle \mathbf{v}, \mathbf{w} \rangle - \beta \langle \mathbf{w}, \mathbf{w} \rangle.$$

This implies that we must have

$$\beta = \frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\langle \mathbf{w}, \mathbf{w} \rangle},$$

and

$$\mathbf{v}_1 = \frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\langle \mathbf{w}, \mathbf{w} \rangle}\mathbf{w}, \qquad \mathbf{v}_2 = \mathbf{v} - \frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\langle \mathbf{w}, \mathbf{w} \rangle}\mathbf{w}.$$

It is easy to check that $\mathbf{v}_1$ and $\mathbf{v}_2$ given by these formulas indeed satisfy the requirements that $\mathbf{v}_1$ is parallel to $\mathbf{w}$ and $\mathbf{v}_2$ is orthogonal to $\mathbf{w}$. This establishes the existence and uniqueness of $\mathbf{v}_1$ and $\mathbf{v}_2$.
Now for any real number α,

$$\mathbf{v} - \alpha\mathbf{w} = \mathbf{v} - \mathbf{v}_1 + (\beta - \alpha)\mathbf{w}.$$

Since $\mathbf{v} - \mathbf{v}_1 = \mathbf{v}_2$ is orthogonal to $(\beta - \alpha)\mathbf{w}$, the generalized Pythagoras theorem implies that

$$\|\mathbf{v} - \alpha\mathbf{w}\|^2 = \|\mathbf{v} - \mathbf{v}_1\|^2 + \|(\beta - \alpha)\mathbf{w}\|^2 \geq \|\mathbf{v} - \mathbf{v}_1\|^2.$$

This proves that

$$\|\mathbf{v} - \alpha\mathbf{w}\| \geq \|\mathbf{v} - \mathbf{v}_1\|.$$

The equality holds if and only if

$$\|(\beta - \alpha)\mathbf{w}\| = |\alpha - \beta|\|\mathbf{w}\| = 0.$$

Since $\|\mathbf{w}\| \neq 0$, we must have $\alpha = \beta$.

The vector $\mathbf{v}_1$ in this theorem is called the projection of $\mathbf{v}$ onto the subspace spanned by $\mathbf{w}$.

There is a more general projection theorem where the subspace W spanned by $\mathbf{w}$ is replaced by a general subspace. We say that a vector $\mathbf{v}$ is orthogonal to the subspace W if it is orthogonal to each vector $\mathbf{w}$ in W.

Theorem 1.11 General Projection Theorem

Let V be an inner product space, and let W be a finite dimensional subspace of V. If $\mathbf{v}$ is a vector in V, there is a unique way to write $\mathbf{v}$ as a sum of two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, such that $\mathbf{v}_1$ is in W and $\mathbf{v}_2$ is orthogonal to W. The vector $\mathbf{v}_1$ is denoted by $\text{proj}_W\mathbf{v}$. For any $\mathbf{w} \in W$,

$$\|\mathbf{v} - \mathbf{w}\| \geq \|\mathbf{v} - \text{proj}_W\mathbf{v}\|,$$

and the equality holds if and only if $\mathbf{w} = \text{proj}_W\mathbf{v}$.

Sketch of Proof

If W is a k- dimensional vector space, it has a basis consists of k linearly independent vectors $\mathbf{w}_1, \ldots, \mathbf{w}_k$. Since the vector $\mathbf{v}_1$ is in W, there are constants $c_1, \ldots, c_k$ such that

$$\mathbf{v}_1 = c_1\mathbf{w}_1 + \cdots + c_k\mathbf{w}_k.$$

The condition $\mathbf{v}_2 = \mathbf{v} - \mathbf{v}_1$ is orthogonal to W gives rise to k equations

$$\begin{aligned} c_1\langle \mathbf{w}_1, \mathbf{w}_1\rangle + \cdots + c_k\langle \mathbf{w}_k, \mathbf{w}_1\rangle &= \langle \mathbf{v}, \mathbf{w}_1\rangle, \\ &\vdots \\ c_1\langle \mathbf{w}_1, \mathbf{w}_k\rangle + \cdots + c_k\langle \mathbf{w}_k, \mathbf{w}_k\rangle &= \langle \mathbf{v}, \mathbf{w}_k\rangle. \end{aligned} \tag{1.2}$$

Using the fact that $\mathbf{w}_1, \ldots, \mathbf{w}_k$ are linearly independent, one can show that the $k \times k$ matrix

$$A = \begin{bmatrix} \langle \mathbf{w}_1, \mathbf{w}_1\rangle & \cdots & \langle \mathbf{w}_k, \mathbf{w}_1\rangle \\ \vdots & \ddots & \vdots \\ \langle \mathbf{w}_1, \mathbf{w}_k\rangle & \cdots & \langle \mathbf{w}_k, \mathbf{w}_k\rangle \end{bmatrix}$$

is invertible. This shows that there is a unique $\mathbf{c} = (c_1, \ldots, c_k)$ satisfying the linear system (1.2).

If V is an inner product space, a basis that consists of mutually orthogonal vectors are of special interest.

Definition 1.14 Orthogonal Set and Orthonormal Set

Let V be an inner product space. A subset of vectors $S = \{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is called an orthogonal set if any two distinct vectors $\mathbf{u}_i$ and $\mathbf{u}_j$ in S are orthogonal. Namely,

$$\langle \mathbf{u}_i, \mathbf{u}_j\rangle = 0 \qquad \text{if } i \neq j.$$

S is called an orthonormal set if it is an orthogonal set of unit vectors. Namely,

$$\langle \mathbf{u}_i, \mathbf{u}_j\rangle = \begin{cases} 0 & \text{if } i \neq j, \\ 1 & \text{if } i = j \end{cases}.$$

If $S = \{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is an orthogonal set of nonzero vectors, it is a linearly independent set of vectors. One can construct an orthonormal set by normalizing each vector in the set. There is a standard algorithm, known as the Gram-Schmidt process, which can turn any linearly independent set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ into

an orthogonal set $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ of nonzero vectors. We start by the following lemma.

Lemma 1.12

Let V be an inner product space, and let $S = \{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ be an orthogonal set of nonzero vectors in V that spans the subspace W. Given any vector $\mathbf{v}$ in V,

$$\text{proj}_W \mathbf{v} = \sum_{i=1}^{k} \frac{\langle \mathbf{v}, \mathbf{u}_i \rangle}{\langle \mathbf{u}_i, \mathbf{u}_i \rangle} \mathbf{u}_i.$$

Proof

By the general projection theorem, $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$, where $\mathbf{v}_1 = \text{proj}_W \mathbf{v}$ is in W and $\mathbf{v}_2$ is orthogonal to W. Since S is a basis for W, there exist scalars $c_1, c_2, \ldots, c_k$ such that $\mathbf{v}_1 = c_1\mathbf{u}_1 + \cdots + c_k\mathbf{u}_k$. Therefore,

$$\mathbf{v} = c_1\mathbf{u}_1 + \cdots + c_k\mathbf{u}_k + \mathbf{v}_2.$$

Since S is an orthogonal set of vectors and $\mathbf{v}_2$ is orthogonal to each $\mathbf{u}_i$, we find that for $1 \leq i \leq k$,

$$\langle \mathbf{v}, \mathbf{u}_i \rangle = c_i \langle \mathbf{u}_i, \mathbf{u}_i \rangle.$$

This proves the lemma.

Theorem 1.13 Gram-Schmidt Process

Let V be an inner product space, and assume that $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is a linearly independent set of vectors in V. Define the vectors $\mathbf{u}_1, \ldots, \mathbf{u}_k$ inductively by $\mathbf{u}_1 = \mathbf{v}_1$, and for $2 \leq j \leq k$,

$$\mathbf{u}_j = \mathbf{v}_j - \sum_{i=1}^{j-1} \frac{\langle \mathbf{v}_j, \mathbf{u}_i \rangle}{\langle \mathbf{u}_i, \mathbf{u}_i \rangle} \mathbf{u}_i.$$

Then $S' = \{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is a nonzero set of orthogonal vectors. Moreover, for each $1 \leq j \leq k$, the set $\{\mathbf{u}_i \,|\, 1 \leq i \leq j\}$ spans the same subspace as the set $\{\mathbf{v}_i \,|\, 1 \leq i \leq j\}$.

Sketch of Proof

For $1 \leq j \leq k$, let W_j be the subspace spanned by the set $\{\mathbf{v}_i \,|\, 1 \leq i \leq j\}$. The vectors $\mathbf{u}_1, \ldots, \mathbf{u}_k$ are constructed by letting $\mathbf{u}_1 = \mathbf{v}_1$, and for $2 \leq j \leq k$,

$$\mathbf{u}_j = \mathbf{v}_j - \text{proj}_{W_{j-1}}\mathbf{v}_j.$$

Since $\{\mathbf{v}_1, \ldots, \mathbf{v}_j\}$ is a linearly independent set, $\mathbf{u}_j \neq \mathbf{0}$. Using induction, one can show that $\text{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_j\} = \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_j\}$. By projection theorem, $\mathbf{u}_j$ is orthogonal to W_{j-1}. Hence, it is orthogonal to $\mathbf{u}_1, \ldots, \mathbf{u}_{j-1}$. This proves the theorem.

A mapping between two vector spaces that respect the linear structures is called a linear transformation.

Definition 1.15 Linear Transformation

Let V and W be real vector spaces. A mapping $T : V \to W$ is called a linear transformation provided that for any $\mathbf{v}_1, \ldots, \mathbf{v}_k$ in V, for any real numbers $c_1, \ldots, c_k$,

$$T(c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k) = c_1T(\mathbf{v}_1) + \cdots + c_kT(\mathbf{v}_k).$$

Linear transformations play important roles in multivariable analysis. In the following, we first define a special class of linear transformations associated to special projections.

For $1 \leq i \leq n$, let $\mathbb{L}_i$ be the subspace of $\mathbb{R}^n$ spanned by the unit vector $\mathbf{e}_i$. For the point $\mathbf{x} = (x_1, \ldots, x_n)$,

$$\text{proj}_{\mathbb{L}_i}\mathbf{x} = x_i\mathbf{e}_i.$$

The number x_i is the i^{th}-component of $\mathbf{x}$. It will play important roles later. The mapping from $\mathbf{x}$ to x_i is a function from $\mathbb{R}^n$ to $\mathbb{R}$.

Definition 1.16 Projection Functions

For $1 \leq i \leq n$, the i^{th}-projection function on $\mathbb{R}^n$ is the function $\pi_i : \mathbb{R}^n \to \mathbb{R}$ defined by

$$\pi_i(\mathbf{x}) = \pi_i(x_1, \ldots, x_n) = x_i.$$

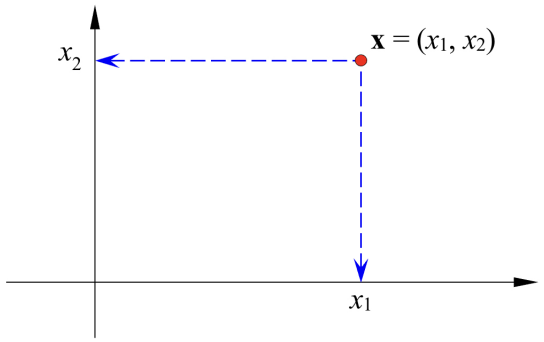


Figure 1.5: The projection functions.

The following is obvious.

Proposition 1.14

For $1 \leq i \leq n$, the i^{th}-projection function on $\mathbb{R}^n$ is a linear transformation. Namely, for any $\mathbf{x}_1, \ldots, \mathbf{x}_k$ in $\mathbb{R}^n$, and any real numbers $c_1, \ldots, c_k$,

$$\pi_i\left(c_1\mathbf{x}_1 + \cdots + c_k\mathbf{x}_k\right) = c_1\pi_i(\mathbf{x}_1) + \cdots + c_k\pi_i(\mathbf{x}_k).$$

The following is a useful inequality.

Proposition 1.15

Let $\mathbf{x}$ be a vector in $\mathbb{R}^n$. Then

$$|\pi_i(\mathbf{x})| \leq \|\mathbf{x}\|.$$

At the end of this section, let us introduce the concept of hyperplanes.

Definition 1.17 Hyperplanes

In $\mathbb{R}^n$, a hyperplane is a translate of a subspace of dimension $n-1$. In other words, $\mathbb{H}$ is a hyperplane if there is a point $\mathbf{x}_0$ in $\mathbb{R}^n$, and $n-1$ linearly independent vectors $\mathbf{v}_1$, $\mathbf{v}_2$, ..., $\mathbf{v}_{n-1}$ such that $\mathbb{H}$ contains all points $\mathbf{x}$ of the form

$$\mathbf{x} = \mathbf{x}_0 + t_1\mathbf{v}_1 + \cdots + t_{n-1}\mathbf{v}_{n-1}, \qquad (t-1, \ldots, t_{n-1}) \in \mathbb{R}^{n-1}.$$

A hyperplane in $\mathbb{R}^1$ is a point. A hyperplane in $\mathbb{R}^2$ is a line. A hyperplane in $\mathbb{R}^3$ is a plane.

Definition 1.18 Normal Vectors

Let $\mathbf{v}_1$, $\mathbf{v}_2$, ..., $\mathbf{v}_{n-1}$ be linearly independent vectors in $\mathbb{R}^n$, and let $\mathbb{H}$ be the hyperplane

$$\mathbb{H} = \left\{ \mathbf{x}_0 + t_1\mathbf{v}_1 + \cdots + t_{n-1}\mathbf{v}_{n-1} \,|\, (t_1, \ldots, t_{n-1}) \in \mathbb{R}^{n-1} \right\}.$$

A nonzero vector $\mathbf{n}$ that is orthogonal to all the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{n-1}$ is called a normal vector to the hyperplane. If $\mathbf{x}_1$ and $\mathbf{x}_2$ are two points on $\mathbb{H}$, then $\mathbf{n}$ is orthogonal to the vector $\mathbf{v} = \mathbf{x}_2 - \mathbf{x}_1$. Any two normal vectors of a hyperplane are scalar multiples of each other.

Proposition 1.16

If $\mathbb{H}$ is a hyperplane with normal vector $\mathbf{n} = (a_1, a_2, \ldots, a_n)$, and $\mathbf{x}_0 = (u_1, u_2, \ldots, u_n)$ is a point on $\mathbb{H}$, then the equation of $\mathbb{H}$ is given by

$$a_1(x_1 - u_1) + a_2(x_2 - u_2) + \cdots + a_n(x_n - u_n) = \mathbf{n} \cdot (\mathbf{x} - \mathbf{x}_0) = 0.$$

Conversely, any equation of the form

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = b$$

is the equation of a hyperplane with normal vector $\mathbf{n} = (a_1, a_2, \ldots, a_n)$.

Example 1.10

Given $1 \leq i \leq n$, the equation $x_i = c$ is a hyperplane with normal vector $\mathbf{e}_i$. It is a hyperplane parallel to the coordinate plane $x_i = 0$, and perpendicular to the x_i-axis.

Exercises 1.1

Question 1

Let V be an inner product space. If $\mathbf{u}$ and $\mathbf{v}$ are vectors in V, show that

$$\big|\|\mathbf{u}\| - \|\mathbf{v}\|\big| \leq \|\mathbf{u} - \mathbf{v}\|.$$

Question 2

Let V be an inner product space. If $\mathbf{u}$ and $\mathbf{v}$ are orthogonal vectors in V, show that

$$\|\mathbf{u} + \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2.$$

Question 3

Let V be an inner product space, and let $\mathbf{u}$ and $\mathbf{v}$ be vectors in V. Show that

$$\langle \mathbf{u}, \mathbf{v} \rangle = \frac{\|\mathbf{u} + \mathbf{v}\|^2 - \|\mathbf{u} - \mathbf{v}\|^2}{4}.$$

Question 4

Let V be an inner product space, and let $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ be an orthonormal set of vectors in V. For any real numbers $\alpha_1, \ldots, \alpha_k$, show that

$$\|\alpha_1 \mathbf{u}_1 + \cdots + \alpha_k \mathbf{u}_k\|^2 = \alpha_1^2 + \cdots + \alpha_k^2.$$

Question 5

Let $x_1, x_2, \ldots, x_n$ be real numbers. Show that

(a) $\sqrt{x_1^2 + x_2^2 \cdots + x_n^2} \leq |x_1| + |x_2| + \cdots + |x_n|$;

(b) $|x_1 + x_2 + \cdots + x_n| \leq \sqrt{n}\sqrt{x_1^2 + x_2^2 \cdots + x_n^2}$.

1.2 Convergence of Sequences in $\mathbb{R}^n$

A point in the Euclidean space $\mathbb{R}^n$ is denoted by $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. When $n = 1$, we just denote it by x. When $n = 2$ and $n = 3$, it is customary to denote a point in $\mathbb{R}^2$ and $\mathbb{R}^3$ by (x, y) and (x, y, z) respectively.

The Euclidean inner product between the vectors $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ is

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{n} x_i y_i.$$

The norm of $\mathbf{x}$ is

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = \sqrt{\sum_{i=1}^{n} x_i^2},$$

while the distance between $\mathbf{x}$ and $\mathbf{y}$ is

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

A sequence in $\mathbb{R}^n$ is a function $f : \mathbb{Z}^+ \to \mathbb{R}^n$. For $k \in \mathbb{Z}^+$, let $\mathbf{a}_k = f(k)$. Then we can also denote the sequence by $\{\mathbf{a}_k\}_{k=1}^{\infty}$, or simply as $\{\mathbf{a}_k\}$.

Example 1.11

The sequence $\left\{ \left(\frac{k}{k+1}, \frac{2k+3}{k} \right) \right\}$ is a sequence in $\mathbb{R}^2$ with

$$\mathbf{a}_k = \left(\frac{k}{k+1}, \frac{2k+3}{k} \right).$$

In volume I, we have seen that a sequence of real numbers $\{a_k\}_{k=1}^{\infty}$ is said to converge to a real number a provided that for any $\varepsilon > 0$, there is a positive integer K such that

$$|a_k - a| < \varepsilon \qquad \text{for all } k \geq K.$$

Notice that $|a_k - a|$ is the distance between a_k and a. To define the convergence of a sequence in $\mathbb{R}^n$, we use the Euclidean distance.

Definition 1.19 Convergence of Sequences

A sequence $\{\mathbf{a}_k\}$ in $\mathbb{R}^n$ is said to converge to the point $\mathbf{a}$ in $\mathbb{R}^n$ provided that for any $\varepsilon > 0$, there is a positive integer K so that for all $k \geq K$,

$$\|\mathbf{a}_k - \mathbf{a}\| = d(\mathbf{a}_k, \mathbf{a}) < \varepsilon.$$

If $\{\mathbf{a}_k\}$ is a sequence that converges to a point $\mathbf{a}$, we say that the sequence $\{\mathbf{a}_k\}$ is convergent. A sequence that does not converge to any point in $\mathbb{R}^n$ is said to be divergent.

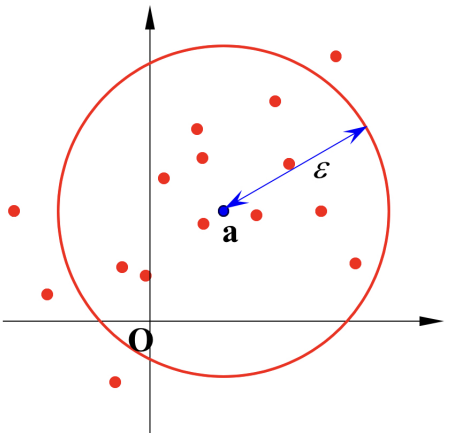


Figure 1.6: The convergence of a sequence.

As in the $n = 1$ case, we have the following.

Proposition 1.17

A sequence in $\mathbb{R}^n$ cannot converge to two different points.

Definition 1.20 Limit of a Sequence

If $\{\mathbf{a}_k\}$ is a sequence in $\mathbb{R}^n$ that converges to the point $\mathbf{a}$, we call $\mathbf{a}$ the limit of the sequence. This can be expressed as

$$\lim_{k\to\infty} \mathbf{a}_k = \mathbf{a}.$$

The following is easy to establish.

Proposition 1.18

Let $\{\mathbf{a}_k\}$ be a sequence in $\mathbb{R}^n$. Then $\{\mathbf{a}_k\}$ converges to $\mathbf{a}$ if and only if

$$\lim_{k\to\infty} \|\mathbf{a}_k - \mathbf{a}\| = 0.$$

Proof

By definition, the sequence $\{\mathbf{a}_k\}$ is convergent if and only if for any $\varepsilon > 0$, there is a positive integer K so that for all $k \geq K$, $\|\mathbf{a}_k - \mathbf{a}\| < \varepsilon$. This is the definition of $\lim_{k\to\infty} \|\mathbf{a}_k - \mathbf{a}\| = 0$.

As in the $n = 1$ case, $\{\mathbf{a}_{k_j}\}_{j=1}^{\infty}$ is a subsequence of $\{\mathbf{a}_k\}$ if $k_1, k_2, k_3, \ldots$ is a strictly increasing subsequence of positive integers.

Corollary 1.19

If $\{\mathbf{a}_k\}$ is a sequence in $\mathbb{R}^n$ that converges to the point $\mathbf{a}$, then any subsequence of $\{\mathbf{a}_k\}$ also converges to $\mathbf{a}$.

Example 1.12

Let us investigate the convergence of the sequence $\{\mathbf{a}_k\}$ in $\mathbb{R}^2$ with

$$\mathbf{a}_k = \left(\frac{k}{k+1}, \frac{2k+3}{k}\right)$$

that is defined in Example 1.11. Notice that

$$\lim_{k\to\infty} \pi_1(\mathbf{a}_k) = \lim_{k\to\infty} \frac{k}{k+1} = 1,$$

$$\lim_{k\to\infty} \pi_2(\mathbf{a}_k) = \lim_{k\to\infty} \frac{2k+3}{k} = 2.$$

It is natural for us to speculate that the sequence $\{\mathbf{a}_k\}$ converges to the point $\mathbf{a} = (1, 2)$.

For $k \in \mathbb{Z}^+$,

$$\mathbf{a}_k - \mathbf{a} = \left(-\frac{1}{k+1}, \frac{3}{k} \right).$$

Thus,

$$\|\mathbf{a}_k - \mathbf{a}\| = \sqrt{\frac{1}{(k+1)^2} + \frac{9}{k^2}}.$$

By squeeze theorem,

$$\lim_{k \to \infty} \|\mathbf{a}_k - \mathbf{a}\| = 0.$$

This proves that the sequence $\{\mathbf{a}_k\}$ indeed converges to the point $\mathbf{a} = (1, 2)$.

In the example above, we guess the limit of the sequence by looking at each components of the sequence. This in fact works for any sequences.

Theorem 1.20 Componentwise Convergence of Sequences

A sequence $\{\mathbf{a}_k\}$ in $\mathbb{R}^n$ converges to the point $\mathbf{a}$ if and only if for each $1 \leq i \leq n$, the sequence $\{\pi_i(\mathbf{a}_k)\}$ converges to the point $\{\pi_i(\mathbf{a})\}$.

Proof

Given $1 \leq i \leq n$,

$$\pi_i(\mathbf{a}_k) - \pi_i(\mathbf{a}) = \pi_i(\mathbf{a}_k - \mathbf{a}).$$

Thus,

$$|\pi_i(\mathbf{a}_k) - \pi_i(\mathbf{a})| = |\pi_i(\mathbf{a}_k - \mathbf{a})| \leq \|\mathbf{a}_k - \mathbf{a}\|.$$

If the sequence $\{\mathbf{a}_k\}$ converges to the point $\mathbf{a}$, then

$$\lim_{k \to \infty} \|\mathbf{a}_k - \mathbf{a}\| = 0.$$

By squeeze theorem,

$$\lim_{k \to \infty} |\pi_i(\mathbf{a}_k) - \pi_i(\mathbf{a})| = 0.$$

This proves that the sequence $\{\pi_i(\mathbf{a}_k)\}$ converges to the point $\{\pi_i(\mathbf{a})\}$.

Conversely, assume that for each $1 \leq i \leq n$, the sequence $\{\pi_i(\mathbf{a}_k)\}$ converges to the point $\{\pi_i(\mathbf{a})\}$. Then

$$\lim_{k\to\infty} |\pi_i(\mathbf{a}_k) - \pi_i(\mathbf{a})| = 0 \qquad \text{for } 1 \leq i \leq n.$$

Since

$$\|\mathbf{a}_k - \mathbf{a}\| \leq \sum_{i=1}^{n} |\pi_i(\mathbf{a}_k - \mathbf{a})|\,,$$

squeeze theorem implies that

$$\lim_{k\to\infty} \|\mathbf{a}_k - \mathbf{a}\| = 0.$$

This proves that the sequence $\{\mathbf{a}_k\}$ converges to the point $\mathbf{a}$.

Theorem 1.20 reduces the investigations of convergence of sequences in $\mathbb{R}^n$ to sequences in $\mathbb{R}$. Let us look at a few examples.

Example 1.13

Find the following limit.

$$\lim_{k\to\infty} \left(\frac{2^k + 1}{3^k}, \left(1 + \frac{1}{k}\right)^k, \frac{k}{\sqrt{k^2+1}} \right).$$

Solution

We compute the limit componentwise.

$$\lim_{k\to\infty} \frac{2^k + 1}{3^k} = \lim_{k\to\infty} \left[\left(\frac{2}{3}\right)^k + \left(\frac{1}{3}\right)^k \right] = 0 + 0 = 0,$$

$$\lim_{k\to\infty} \left(1 + \frac{1}{k}\right)^k = e,$$

$$\lim_{k\to\infty} \frac{k}{\sqrt{k^2+1}} = \lim_{k\to\infty} \frac{k}{k\sqrt{1 + \dfrac{1}{k^2}}} = 1.$$

Hence,

$$\lim_{k\to\infty}\left(\frac{2^k+1}{3^k}, \left(1+\frac{1}{k}\right)^k, \frac{k}{\sqrt{k^2+1}}\right) = (0, e, 1).$$

Example 1.14

Let $\{\mathbf{a}_k\}$ be the sequence with

$$\mathbf{a}_k = \left((-1)^k, \frac{(-1)^k}{k}\right).$$

Is the sequence convergent? Justify your answer.

Solution

The sequence $\{\pi_1(\mathbf{a}_k)\}$ is the sequence $\{(-1)^k\}$, which is divergent. Hence, the sequence $\{\mathbf{a}_k\}$ is divergent.

Using the componentwise convergence theorem, it is easy to establish the following.

Proposition 1.21 Linearity

Let $\{\mathbf{a}_k\}$ and $\{\mathbf{b}_k\}$ be sequences in $\mathbb{R}^n$ that converges to $\mathbf{a}$ and $\mathbf{b}$ respectively. For any real numbers α and β, the sequence $\{\alpha\mathbf{a}_k + \beta\mathbf{b}_k\}$ converges to $\alpha\mathbf{a} + \beta\mathbf{b}$. Namely,

$$\lim_{k\to\infty}(\alpha\mathbf{a}_k + \beta\mathbf{b}_k) = \alpha\mathbf{a} + \beta\mathbf{b}.$$

Example 1.15

If $\{\mathbf{a}_k\}$ is a sequence in $\mathbb{R}^n$ that converges to $\mathbf{a}$, show that

$$\lim_{k\to\infty}\|\mathbf{a}_k\| = \|\mathbf{a}\|.$$

Solution

Notice that

$$\|\mathbf{a}_k\| = \sqrt{\pi_1(\mathbf{a}_k)^2 + \cdots + \pi_n(\mathbf{a}_k)^2}.$$

For $1 \leq i \leq n$,

$$\lim_{k\to\infty} \pi_i(\mathbf{a}_k) = \pi_i(\mathbf{a}).$$

Using limit laws for sequences in $\mathbb{R}$, we have

$$\lim_{k\to\infty} \left(\pi_1(\mathbf{a}_k)^2 + \cdots + \pi_n(\mathbf{a}_k)^2\right) = \pi_1(\mathbf{a})^2 + \cdots + \pi_n(\mathbf{a})^2.$$

Using the fact that square root function is continuous, we find that

$$\begin{aligned}\lim_{k\to\infty} \|\mathbf{a}_k\| &= \lim_{k\to\infty} \sqrt{\pi_1(\mathbf{a}_k)^2 + \cdots + \pi_n(\mathbf{a}_k)^2} \\ &= \sqrt{\pi_1(\mathbf{a})^2 + \cdots + \pi_n(\mathbf{a})^2} = \|\mathbf{a}\|.\end{aligned}$$

There is also a Cauchy criterion for convergence of sequences in $\mathbb{R}^n$.

Definition 1.21 Cauchy Sequences

A sequence $\{\mathbf{a}_k\}$ in $\mathbb{R}^n$ is a Cauchy sequence if for every $\varepsilon > 0$, there is a positive integer K such that for all $l \geq k \geq K$,

$$\|\mathbf{a}_l - \mathbf{a}_k\| < \varepsilon.$$

Theorem 1.22 Cauchy Criterion

A sequence $\{\mathbf{a}_k\}$ in $\mathbb{R}^n$ is convergent if and only if it is a Cauchy sequence.

Similar to the $n = 1$ case, the Cauchy criterion allows us to determine whether a sequence in $\mathbb{R}^n$ is convergent without having to guess what is the limit first.

Proof

Assume that the sequence $\{\mathbf{a}_k\}$ converges to $\mathbf{a}$. Given $\varepsilon > 0$, there is a positive integer K such that for all $k \geq K$, $\|\mathbf{a}_k - \mathbf{a}\| < \varepsilon/2$. Then for all $l \geq k \geq K$,

$$\|\mathbf{a}_l - \mathbf{a}_k\| \leq \|\mathbf{a}_l - \mathbf{a}\| + \|\mathbf{a}_k - \mathbf{a}\| < \varepsilon.$$

This proves that $\{\mathbf{a}_k\}$ is a Cauchy sequence.
Conversely, assume that $\{\mathbf{a}_k\}$ is a Cauchy sequence. Given $\varepsilon > 0$, there is a positive integer K such that for all $l \geq k \geq K$,

$$\|\mathbf{a}_l - \mathbf{a}_k\| < \varepsilon.$$

For each $1 \leq i \leq n$,

$$|\pi_i(\mathbf{a}_l) - \pi_i(\mathbf{a}_k)| = |\pi_i(\mathbf{a}_l - \mathbf{a}_k)| \leq \|\mathbf{a}_l - \mathbf{a}_k\|.$$

Hence, $\{\pi_i(\mathbf{a}_k)\}$ is a Cauchy sequence in $\mathbb{R}$. Therefore, it is convergent. By componentwise convergence theorem, the sequence $\{\mathbf{a}_k\}$ is convergent.

Exercises 1.2

Question 1

Show that a sequence in $\mathbb{R}^n$ cannot converge to two different points.

Question 2

Find the limit of the sequence $\{\mathbf{a}_k\}$, where

$$\mathbf{a}_k = \left(\frac{2k+1}{k+3}, \frac{\sqrt{2k^2+k}}{k}, \left(1+\frac{2}{k}\right)^k \right).$$

Question 3

Let $\{\mathbf{a}_k\}$ be the sequence with

$$\mathbf{a}_k = \left(\frac{1+(-1)^{k-1}k}{1+k}, \frac{1}{2^k} \right).$$

Determine whether the sequence is convergent.

Question 4

Let $\{\mathbf{a}_k\}$ be the sequence with

$$\mathbf{a}_k = \left(\frac{k}{1+k}, \frac{k}{\sqrt{k}+1} \right).$$

Determine whether the sequence is convergent.

Question 5

Let $\{\mathbf{a}_k\}$ and $\{\mathbf{b}_k\}$ be sequences in $\mathbb{R}^n$ that converges to $\mathbf{a}$ and $\mathbf{b}$ respectively. Show that

$$\lim_{k\to\infty} \langle \mathbf{a}_k, \mathbf{b}_k \rangle = \langle \mathbf{a}, \mathbf{b} \rangle.$$

Here $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x} \cdot \mathbf{y}$ is the standard inner product on $\mathbb{R}^n$.

Question 6

Suppose that $\{\mathbf{a}_k\}$ is a sequence in $\mathbb{R}^n$ that converges to $\mathbf{a}$, and $\{c_k\}$ is a sequence of real numbers that converges to c, show that

$$\lim_{k\to\infty} c_k\mathbf{a}_k = c\mathbf{a}.$$

Question 7

Suppose that $\{\mathbf{a}_k\}$ is a sequence of nonzero vectors in $\mathbb{R}^n$ that converges to $\mathbf{a}$ and $\mathbf{a} \neq \mathbf{0}$, show that

$$\lim_{k\to\infty} \frac{\mathbf{a}_k}{\|\mathbf{a}_k\|} = \frac{\mathbf{a}}{\|\mathbf{a}\|}.$$

Question 8

Let $\{\mathbf{a}_k\}$ and $\{\mathbf{b}_k\}$ be sequences in $\mathbb{R}^n$. If $\{\mathbf{a}_k\}$ is convergent and $\{\mathbf{b}_k\}$ is divergent, show that the sequence $\{\mathbf{a}_k + \mathbf{b}_k\}$ is divergent.

Question 9

Suppose that $\{\mathbf{a}_k\}$ is a sequence in $\mathbb{R}^n$ that converges to $\mathbf{a}$. If $r = \|\mathbf{a}\| \neq 0$, show that there is a positive integer K such that

$$\|\mathbf{a}_k\| > \frac{r}{2} \qquad \text{for all } k \geq K.$$

Question 10

Let $\{\mathbf{a}_k\}$ be a sequence in $\mathbb{R}^n$ and let $\mathbf{b}$ be a point in $\mathbb{R}^n$. Assume that the sequence $\{\mathbf{a}_k\}$ does not converge to $\mathbf{b}$. Show that there is an $\varepsilon > 0$ and a subsequence $\{\mathbf{a}_{k_j}\}$ of $\{\mathbf{a}_k\}$ such that

$$\|\mathbf{a}_{k_j} - \mathbf{b}\| \geq \varepsilon \qquad \text{for all } j \in \mathbb{Z}^+.$$

1.3 Open Sets and Closed Sets

In volume I, we call an interval of the form (a, b) an *open* interval. Given a point x in $\mathbb{R}$, a neighbourhood of x is an open interval (a, b) that contains x. Given a subset S of $\mathbb{R}$, we say that x is an interior point of S if there is a neighboirhood of x that is contained in S. We say that S is closed in $\mathbb{R}$ provided that if $\{a_k\}$ is a sequence of points in S that converges to a, then a is also in S. These describe the topology of $\mathbb{R}$. It is relatively simple.

For $n \geq 2$, the topological features of $\mathbb{R}^n$ are much more complicated.

An open interval (a, b) in $\mathbb{R}$ can be described as a set of the form

$$B = \{x \in \mathbb{R} \,|\, |x - x_0| < r\}\,,$$

where $x_0 = \dfrac{a+b}{2}$ and $r = \dfrac{b-a}{2}$.

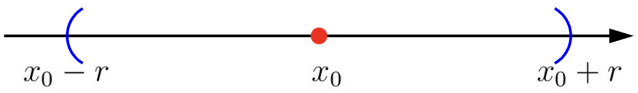


Figure 1.7: An open interval.

Generalizing this, we define open balls in $\mathbb{R}^n$.

Definition 1.22 Open Balls

Given $\mathbf{x}_0$ in $\mathbb{R}^n$ and $r > 0$, an open ball $B(\mathbf{x}_0, r)$ of radius r with center at $\mathbf{x}_0$ is a subset of $\mathbb{R}^n$ of the form

$$B(\mathbf{x}_0, r) = \{\mathbf{x} \in \mathbb{R}^n \,|\, \|\mathbf{x} - \mathbf{x}_0\| < r\}\,.$$

It consists of all points of $\mathbb{R}^n$ whose distance to the center $\mathbf{x}_0$ is less than r.

Obviously, it $0 < r_1 \leq r_2$, then $B(\mathbf{x}_0, r_1) \subset B(\mathbf{x}_0, r_2)$. The following is a useful lemma for balls with different centers.

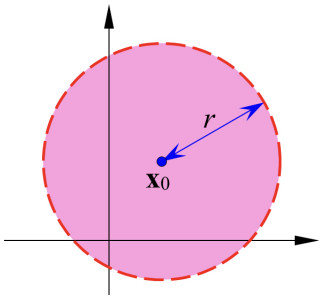


Figure 1.8: An open ball.

Lemma 1.23

Let $\mathbf{x}_1$ be a point in the open ball $B(\mathbf{x}_0, r)$. Then $\|\mathbf{x}_1 - \mathbf{x}_0\| < r$. If r_1 is a positive number satisfying

$$r_1 \leq r - \|\mathbf{x}_1 - \mathbf{x}_0\|,$$

then the open ball $B(\mathbf{x}_1, r_1)$ is contained in the open ball $B(\mathbf{x}_0, r)$.

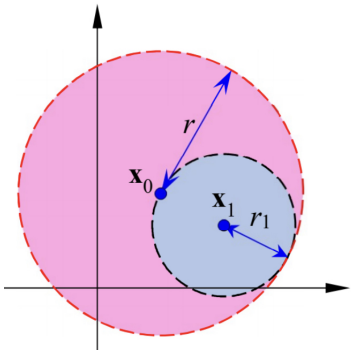


Figure 1.9: An open ball containing another open ball with different center.

Proof

Let $\mathbf{x}$ be a point in $B(\mathbf{x}_1, r_1)$. Then

$$\|\mathbf{x} - \mathbf{x}_1\| < r_1 \leq r - \|\mathbf{x}_1 - \mathbf{x}_0\|.$$

By triangle inequality,

$$\|\mathbf{x} - \mathbf{x}_0\| \leq \|\mathbf{x} - \mathbf{x}_1\| + \|\mathbf{x}_1 - \mathbf{x}_0\| < r.$$

Therefore, $\mathbf{x}$ is a point in $B(\mathbf{x}_0, r)$. This proves the assertion.

Now we define open sets in $\mathbb{R}^n$.

Definition 1.23 Open Sets

Let S be a subset of $\mathbb{R}^n$. We say that S is an open set if for each $\mathbf{x} \in S$, there is a ball $B(\mathbf{x}, r)$ centered at $\mathbf{x}$ that is contained in S.

The following example justifies that an open interval of the form (a, b) is an open set.

Example 1.16

Let S to be the open interval $S = (a, b)$ in $\mathbb{R}$. If $x \in S$, then $a < x < b$. Hence, $x - a$ and $b - x$ are positive. Let $r = \min\{x - a, b - x\}$. Then $r > 0$, $r \leq x - a$ and $r \leq b - x$. These imply that $a \leq x - r < x + r \leq b$. Hence, $B(x, r) = (x - r, x + r) \subset (a, b) = S$. This shows that the interval (a, b) is an open set.

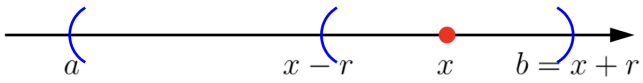


Figure 1.10: The interval (a, b) is an open set.

The following example justifies that an open ball is indeed an open set.

Example 1.17

Let $S = B(\mathbf{x}_0, r)$ be the open ball with center at $\mathbf{x}_0$ and radius $r > 0$ in $\mathbb{R}^n$. Show that S is an open set.

Solution

Given $\mathbf{x} \in S$, $d = \|\mathbf{x} - \mathbf{x}_0\| < r$. Let $r_1 = r - d$. Then $r_1 > 0$. Lemma 1.23 implies that the ball $B(\mathbf{x}, r_1)$ is inside S. Hence, S is an open set.

Example 1.18

As subsets of $\mathbb{R}^n$, $\emptyset$ and $\mathbb{R}^n$ are open sets.

Example 1.19

A one-point set $S = \{\mathbf{a}\}$ in $\mathbb{R}^n$ cannot be open, for there is no $r > 0$ such that $B(\mathbf{a}, r)$ in contained in S.

Let us look at some other examples of open sets.

Definition 1.24 Open Rectangles

A set of the form

$$U = \prod_{i=1}^{n} (a_i, b_i) = (a_1, b_1) \times \cdots \times (a_n, b_n)$$

in $\mathbb{R}^n$, which is a cartesian product of open bounded intervals, in called an open rectangle.

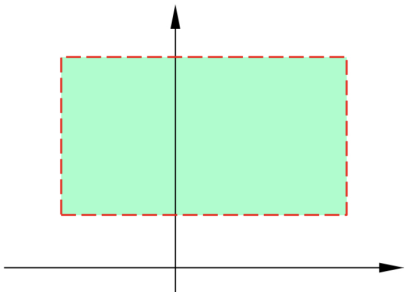


Figure 1.11: A rectangle in $\mathbb{R}^2$.

Example 1.20

Let $U = \prod_{i=1}^{n}(a_i, b_i)$ be an open rectangle in $\mathbb{R}^n$. Show that U is an open set.

Solution

Let $\mathbf{x} = (x_1, \ldots, x_n)$ be a point in U. Then for $1 \leq i \leq n$,

$$r_i = \min\{x_i - a_i, b_i - x_i\} > 0$$

and

$$(x_i - r_i, x_i + r_i) \subset (a_i, b_i).$$

Let $r = \min\{r_1, \ldots, r_n\}$. Then $r > 0$. We claim that $B(\mathbf{x}, r)$ is contained in U.
If $\mathbf{y} \in B(\mathbf{x}, r)$, then $\|\mathbf{y} - \mathbf{x}\| < r$. This implies that

$$|y_i - x_i| \leq \|\mathbf{y} - \mathbf{x}\| < r \leq r_i \qquad \text{for all } 1 \leq i \leq n.$$

Hence,

$$y_i \in (x_i - r_i, x_i + r_i) \subset (a_i, b_i) \qquad \text{for all } 1 \leq i \leq n.$$

This proves that $\mathbf{y} \in U$, and thus, completes the proof that $B(\mathbf{x}, r)$ is contained in U. Therefore, U is an open set.

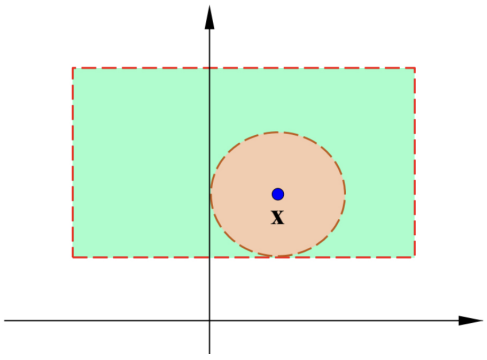


Figure 1.12: An open rectangle is an open set.

Next, we define closed sets. The definition is a straightforward generalization of the $n = 1$ case.

Definition 1.25 Closed Sets

Let S be a subset of $\mathbb{R}^n$. We say that S is closed in $\mathbb{R}^n$ provided that if $\{\mathbf{a}_k\}$ is a sequence of points in S that converges to the point $\mathbf{a}$, the point $\mathbf{a}$ is also in S.

Example 1.21

As subsets of $\mathbb{R}^n$, $\emptyset$ and $\mathbb{R}^n$ are closed sets. Since $\emptyset$ and $\mathbb{R}^n$ are also open, a subset S of $\mathbb{R}^n$ can be both open and closed.

Example 1.22

Let $S = \{\mathbf{a}\}$ be a one-point set in $\mathbb{R}^n$. A sequence $\{\mathbf{a}_k\}$ in S is just the constant sequence where $\mathbf{a}_k = \mathbf{a}$ for all $k \in \mathbb{Z}^+$. Hence, it converges to $\mathbf{a}$ which is in S. Thus, a one-point set S is a closed set.

In volume I, we have proved the following.

Proposition 1.24

Let I be intervals of the form $(-\infty, a]$, $[a, \infty)$ or $[a, b]$. Then I is a closed subset of $\mathbb{R}$.

Definition 1.26 Closed Rectangles

A set of the form

$$R = \prod_{i=1}^{n} [a_i, b_i] = [a_1, b_1] \times \cdots \times [a_n, b_n]$$

in $\mathbb{R}^n$, which is a cartesian product of closed and bounded intervals, is called a closed rectangle.

The following justifies that a closed rectangle is indeed a closed set.

Example 1.23

Let

$$R = \prod_{i=1}^{n} [a_i, b_i] = [a_1, b_1] \times \cdots \times [a_n, b_n]$$

be a closed rectangle in $\mathbb{R}^n$. Show that R is a closed set.

Solution

Let $\{\mathbf{a}_k\}$ be a sequence in R that converges to a point $\mathbf{a}$. For each $1 \leq i \leq n$, $\{\pi_i(\mathbf{a}_k)\}$ is a sequence in $[a_i, b_i]$ that converges to $\pi_i(\mathbf{a})$. Since $[a_i, b_i]$ is a closed set in $\mathbb{R}$, $\pi_i(\mathbf{a}) \in [a_i, b_i]$. Hence, $\mathbf{a}$ is in R. This proves that R is a closed set.

It is not true that a set that is not open is closed.

Example 1.24

Show that an interval of the form $I = (a, b]$ in $\mathbb{R}$ is neither open nor closed.

Solution

If I is open, since b is in I, there is an $r > 0$ such that $(b - r, b + r) = B(b, r) \subset I$. But then $b + r/2$ is a point in $(b - r, b + r)$ but not in $I = (a, b]$, which gives a contradiction. Hence, I is not open.

For $k \in \mathbb{Z}^+$, let

$$a_k = a + \frac{b - a}{k}.$$

Then $\{a_k\}$ is a sequence in I that converges to a, but a is not in I. Hence, I is not closed.

Thus, we have seen that a subset S of $\mathbb{R}^n$ can be both open and closed, and it can also be neither open nor closed.

Let us look at some other examples of closed sets.

Definition 1.27 Closed Balls

Given $\mathbf{x}_0$ in $\mathbb{R}^n$ and $r > 0$, a closed ball of radius r with center at $\mathbf{x}_0$ is a subset of $\mathbb{R}^n$ of the form

$$CB(\mathbf{x}_0, r) = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x} - \mathbf{x}_0\| \leq r\}.$$

It consists of all points of $\mathbb{R}^n$ whose distance to the center $\mathbf{x}_0$ is less than or equal to r.

The following justifies that a closed ball is indeed a closed set.

Example 1.25

Given $\mathbf{x}_0 \in \mathbb{R}^n$ and $r > 0$, show that the closed ball

$$CB(\mathbf{x}_0, r) = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x} - \mathbf{x}_0\| \leq r\}$$

is a closed set.

Solution

Let $\{\mathbf{a}_k\}$ be a sequence in $CB(\mathbf{x}_0, r)$ that converges to the point $\mathbf{a}$. Then

$$\lim_{k\to\infty} \|\mathbf{a}_k - \mathbf{a}\| = 0.$$

For each $k \in \mathbb{Z}^+$, $\|\mathbf{a}_k - \mathbf{x}_0\| \leq r$. By triangle inequality,

$$\|\mathbf{a} - \mathbf{x}_0\| \leq \|\mathbf{a}_k - \mathbf{x}_0\| + \|\mathbf{a}_k - \mathbf{a}\| \leq r + \|\mathbf{a}_k - \mathbf{a}\|.$$

Taking the $k \to \infty$ limit, we find that

$$\|\mathbf{a} - \mathbf{x}_0\| \leq r.$$

Hence, $\mathbf{a}$ is in $CB(\mathbf{x}_0, r)$. This proves that $CB(\mathbf{x}_0, r)$ is a closed set.

The following theorem gives the relation between open and closed sets.

Theorem 1.25

Let S be a subset of $\mathbb{R}^n$ and let $A = \mathbb{R}^n \setminus S$ be its complement in $\mathbb{R}^n$. Then S is open if and only if A is closed.

Proof

Assume that S is open. Let $\{\mathbf{a}_k\}$ be a sequence in A that converges to the point $\mathbf{a}$. We want to show that $\mathbf{a}$ is in A. Assume to the contrary that $\mathbf{a}$ is not in A. Then $\mathbf{a}$ is in S. Since S is open, there is an $r > 0$ such that $B(\mathbf{a}, r)$ is contained in S. Since the sequence $\{\mathbf{a}_k\}$ converges to $\mathbf{a}$, there is a positive integer K such that for all $k \geq K$, $\|\mathbf{a}_k - \mathbf{a}\| < r$. But then this implies that $\mathbf{a}_K \in B(\mathbf{a}, r) \subset S$. This contradicts to $\mathbf{a}_K$ is in $A = \mathbb{R}^n \setminus S$. Hence, we must have $\mathbf{a}$ is in A, which proves that A is closed.
Conversely, assume that A is closed. We want to show that S is open. Assume to the contrary that S is not open. Then there is a point $\mathbf{a}$ in S such that for every $r > 0$, $B(\mathbf{a}, r)$ is not contained in S. For every $k \in \mathbb{Z}^+$, since $B(\mathbf{a}, 1/k)$ is not contained in S, there is a point $\mathbf{a}_k$ in $B(\mathbf{a}, 1/k)$ such that $\mathbf{a}_k$ is not in S. Thus, $\{\mathbf{a}_k\}$ is a sequence in A and

$$\|\mathbf{a}_k - \mathbf{a}\| < \frac{1}{k}.$$

This shows that $\{\mathbf{a}_k\}$ converges to $\mathbf{a}$. Since A is closed, $\mathbf{a}$ is in A, which contradicts to $\mathbf{a}$ is in S. Thus, we must have S is open.

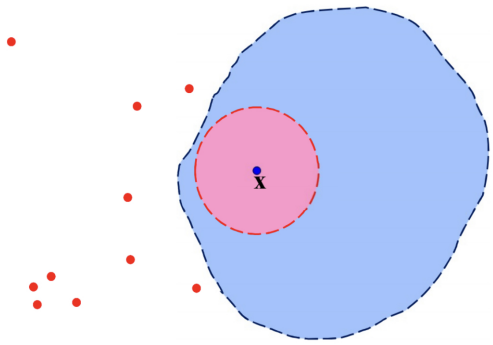


Figure 1.13: A sequence outside an open set cannot converge to a point in the open set.

Next, we consider unions and intersections of sets.

Theorem 1.26

1. Arbitrary union of open sets is open. Namely, if $\{U_\alpha \,|\, \alpha \in J\}$ is a collection of open sets in $\mathbb{R}^n$, then their union $U = \bigcup_{\alpha\in J} U_\alpha$ is also an open set.
2. Finite intersections of open sets is open. Namely, if $V_1, \ldots, V_k$ are open sets in $\mathbb{R}^n$, then their intersection $V = \bigcap_{i=1}^{k} V_i$ is also an open set.

Proof

To prove the first statement, let $\mathbf{x}$ be a point in $U = \bigcup_{\alpha\in J} U_\alpha$. Then there is an $\alpha \in J$ such that $\mathbf{x}$ is in U_α. Since U_α is open, there is an $r > 0$ such that $B(\mathbf{x}, r) \subset U_\alpha \subset U$. Hence, U is open.

For the second statement, let $\mathbf{x}$ be a point in $V = \bigcap_{i=1}^{k} V_i$. Then for each $1 \leq i \leq k$, $\mathbf{x}$ is in the open set V_i. Hence, there is an $r_i > 0$ such that $B(\mathbf{x}, r_i) \subset V_i$. Let $r = \min\{r_1, \ldots, r_k\}$. Then for $1 \leq i \leq k$, $r \leq r_i$ and so $B(\mathbf{x}, r) \subset B(\mathbf{x}, r_i) \subset V_i$. Hence, $B(\mathbf{x}, r) \subset V$. This proves that V is open.

As an application of this theorem, let us show that any open interval in $\mathbb{R}$ is indeed an open set.

Proposition 1.27

Let I be an interval of the form $(-\infty, a)$, (a, ∞) or (a, b). Then I is an open subset of $\mathbb{R}$.

Proof

We have shown in Example 1.16 that if I is an interval of the form (a, b), then I is an open subset of $\mathbb{R}$. Now

$$(a, \infty) = \bigcup_{k=1}^{\infty} (a, a + k)$$

is a union of open sets. Hence, (a, ∞) is open. In the same way, one can show that an interval of the form $(-\infty, a)$ is open.

The next example shows that arbitrary intersections of open sets is not necessary open.

Example 1.26

For $k \in \mathbb{Z}^+$, let U_k be the open set in $\mathbb{R}$ given by

$$U_k = \left(-\frac{1}{k}, \frac{1}{k}\right).$$

Notice that the set

$$U = \bigcap_{k=1}^{\infty} U_k = \{0\}$$

is a one-point set. Hence, it is not open in $\mathbb{R}$.

De Morgan's law in set theory says that if $\{U_\alpha \mid \alpha \in J\}$ is a collection of sets in $\mathbb{R}^n$, then

$$\mathbb{R}^n \setminus \bigcup_{\alpha \in J} U_\alpha = \bigcap_{\alpha \in J} (\mathbb{R}^n \setminus U_\alpha),$$
$$\mathbb{R}^n \setminus \bigcap_{\alpha \in J} U_\alpha = \bigcup_{\alpha \in J} (\mathbb{R}^n \setminus U_\alpha).$$

Thus, we obtain the counterpart of Theorem 1.26 for closed sets.

Theorem 1.28

1. Arbitrary intersection of closed sets is closed. Namely, if $\{A_\alpha \,|\, \alpha \in J\}$ is a collection of closed sets in $\mathbb{R}^n$, then their intersection $A = \bigcap_{\alpha\in J} A_\alpha$ is also a closed set.

2. Finite union of closed sets is closed. Namely, if $C_1, \ldots, C_k$ are closed sets in $\mathbb{R}^n$, then their union $C = \bigcup_{i=1}^{k} C_i$ is also a closed set.

Proof

We prove the first statement. The proof of the second statement is similar. Given that $\{A_\alpha \,|\, \alpha \in J\}$ is a collection of closed sets in $\mathbb{R}^n$, for each $\alpha \in J$, let $U_\alpha = \mathbb{R}^n \setminus A_\alpha$. Then $\{U_\alpha \,|\, \alpha \in J\}$ is a collection of open sets in $\mathbb{R}^n$. By Theorem 1.26, the set $\bigcup_{\alpha\in J} U_\alpha$ is open. By Theorem 1.25, $\mathbb{R}^n \setminus \bigcup_{\alpha\in J} U_\alpha$ is closed. By De Morgan's law,

$$\mathbb{R}^n \setminus \bigcup_{\alpha\in J} U_\alpha = \bigcap_{\alpha\in J} (\mathbb{R}^n \setminus U_\alpha) = \bigcap_{\alpha\in J} A_\alpha.$$

This proves that $\bigcap_{\alpha\in J} A_\alpha$ is a closed set.

The following example says that any finite point set is a closed set.

Example 1.27

Let $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ be a finite point set in $\mathbb{R}^n$. Then $S = \bigcup_{i=1}^{k} \{\mathbf{x}_i\}$ is a finite union of one-point sets. Since one-point set is closed, S is closed.

Exercises 1.3

Question 1

Let A be the subset of $\mathbb{R}^2$ given by

$$A = \{(x, y) \mid x > 0, y > 0\}.$$

Show that A is an open set.

Question 2

Let A be the subset of $\mathbb{R}^2$ given by

$$A = \{(x, y) \mid x \geq 0, y \geq 0\}.$$

Show that A is a closed set.

Question 3

Let A be the subset of $\mathbb{R}^2$ given by

$$A = \{(x, y) \mid x > 0, y \geq 0\}.$$

Is A open? Is A closed? Justify your answers.

Question 4

Let C and U be subsets of $\mathbb{R}^n$. Assume that C is closed and U is open, show that $U \setminus C$ is open and $C \setminus U$ is closed.

Question 5

Let A be a subset of $\mathbb{R}^n$, and let $B = A + \mathbf{u}$ be the translate of A by the vector $\mathbf{u}$.

(a) Show that A is open if and only if B is open.

(b) Show that A is closed if and only if B is closed.

1.4 Interior, Exterior, Boundary and Closure

First, we introduce the interior of a set.

Definition 1.28 Interior

Let S be a subset of $\mathbb{R}^n$. We say that $\mathbf{x} \in \mathbb{R}^n$ is an interior point of S if there exists $r > 0$ such that $B(\mathbf{x}, r) \subset S$. The interior of S, denoted by $\operatorname{int} S$, is defined to be the collection of all the interior points of S.

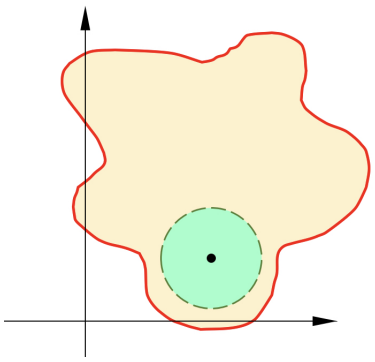


Figure 1.14: The interior point of a set.

The following gives a characterization of the interior of a set.

Theorem 1.29

Let S be a subset of $\mathbb{R}^n$. Then we have the followings.

1. $\operatorname{int} S$ is a subset of S.
2. $\operatorname{int} S$ is an open set.
3. S is an open set if and only if $S = \operatorname{int} S$.
4. If U is an open set that is contained in S, then $U \subset \operatorname{int} S$.

These imply that $\operatorname{int} S$ is the largest open set that is contained in S.

Proof

Let $\mathbf{x}$ be a point in $\operatorname{int} S$. By definition, there exists $r > 0$ such that $B(\mathbf{x}, r) \subset S$. Since $\mathbf{x} \in B(\mathbf{x}, r)$ and $B(\mathbf{x}, r) \subset S$, $\mathbf{x}$ is a point in S. Since we have shown that every point in $\operatorname{int} S$ is in S, $\operatorname{int} S$ is a subset of S.
If $\mathbf{y} \in B(\mathbf{x}, r)$, Lemma 1.23 says that there is an $r' > 0$ such that $B(\mathbf{y}, r') \subset B(\mathbf{x}, r) \subset S$. Hence, $\mathbf{y}$ is also in $\operatorname{int} S$. This proves that $B(\mathbf{x}, r)$ is contained in $\operatorname{int} S$. Since we have shown that for any $\mathbf{x} \in \operatorname{int} S$, there is an $r > 0$ such that $B(\mathbf{x}, r)$ is contained in $\operatorname{int} S$, this shows that $\operatorname{int} S$ is open.
If $S = \operatorname{int} S$, S is open. Conversely, if S is open, for every $\mathbf{x}$ in S, there is an $r > 0$ such that $B(\mathbf{x}, r) \subset S$. Then $\mathbf{x}$ is in $\operatorname{int} S$. Hence, $S \subset \operatorname{int} S$. Since we have shown that $\operatorname{int} S \subset S$ is always true, we conclude that if S is open, $S = \operatorname{int} S$.
If U is a subset of S and U is open, for every $\mathbf{x}$ in U, there is an $r > 0$ such that $B(\mathbf{x}, r) \subset U$. But then $B(\mathbf{x}, r) \subset S$. This shows that $\mathbf{x}$ is in $\operatorname{int} S$. Since every point of U is in $\operatorname{int} S$, this proves that $U \subset \operatorname{int} S$.

Example 1.28

Find the interior of each of the following subsets of $\mathbb{R}$.

(a) $A = (a, b)$

(b) $B = (a, b]$

(c) $C = [a, b]$

(d) $\mathbb{Q}$

Solution

(a) Since A is an open set, $\operatorname{int} A = A = (a, b)$.

(b) Since A is an open set that is contained in B, $A = (a, b)$ is contained in $\operatorname{int} B$. Since $\operatorname{int} B \subset B$, we only left to determine whether b is in $\operatorname{int} B$. The same argument as given in Example 1.24 shows that b is not an interior point of B. Hence, $\operatorname{int} B = A = (a, b)$.

(c) Similar arguments as given in (b) show that $A \subset \operatorname{int} C$, and both a and b are not interior points of C. Hence, $\operatorname{int} C = A = (a, b)$.

(d) For any $x \in \mathbb{R}$ and any $r > 0$, $B(x, r) = (x - r, x + r)$ contains an irrational number. Hence, $B(x, r)$ is not contained in $\mathbb{Q}$. This shows that $\mathbb{Q}$ does not have interior points. Hence, $\operatorname{int} \mathbb{Q} = \emptyset$.

Definition 1.29 Neighbourhoods

Let $\mathbf{x}$ be a point in $\mathbb{R}^n$ and let U be a subset of $\mathbb{R}^n$. We say that U is a neighbourhood of $\mathbf{x}$ if U is an *open* set that contains $\mathbf{x}$.

Notice that this definition is slightly different from the one we use in volume I for the $n = 1$ case.

Neighbourhoods

By definition, if U is a neighbourhood of $\mathbf{x}$, then $\mathbf{x}$ is an interior point of U, and there is an $r > 0$ such that $B(\mathbf{x}, r) \subset U$.

Example 1.29

Consider the point $\mathbf{x} = (1, 2)$ and the sets

$$U = \left\{(x_1, x_2) \,|\, x_1^2 + x_2^2 < 9\right\},$$
$$V = \{(x_1, x_2) \,|\, 0 < x_1 < 2, -1 < x_2 < 3\}$$

in $\mathbb{R}^2$. The sets U and V are neighbourhoods of $\mathbf{x}$.

Next, we introduce the exterior and boundary of a set.

Definition 1.30 Exterior

Let S be a subset of $\mathbb{R}^n$. We say that $\mathbf{x} \in \mathbb{R}^n$ is an exterior point of S if there exists $r > 0$ such that $B(\mathbf{x}, r) \subset \mathbb{R}^n \setminus S$. The exterior of S, denoted by $\operatorname{ext} S$, is defined to be the collection of all the exterior points of S.

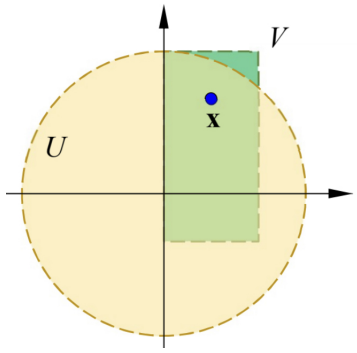


Figure 1.15: The sets U and V are neighbourhoods of the point $\mathbf{x}$.

Definition 1.31 Boundary

Let S be a subset of $\mathbb{R}^n$. We say that $\mathbf{x} \in \mathbb{R}^n$ is a boundary point of S if for every $r > 0$, the ball $B(\mathbf{x}, r)$ intersects both S and $\mathbb{R}^n \setminus S$. The boundary of S, denoted by bd S or ∂S, is defined to be the collection of all the boundary points of S.

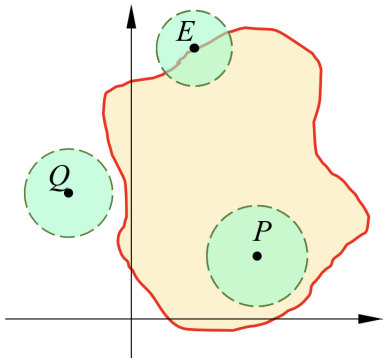


Figure 1.16: P is an interior point, Q is an exterior point, E is a boundary point.

Theorem 1.30

Let S be a subset of $\mathbb{R}^n$. We have the followings.

(a) $\text{ext}\,(S) = \text{int}\,(\mathbb{R}^n \setminus S)$.

(b) $\text{bd}\,(S) = \text{bd}\,(\mathbb{R}^n \setminus S)$.

(c) $\text{int}\,S$, $\text{ext}\,S$ and $\text{bd}\,S$ are mutually disjoint sets.

(d) $\mathbb{R}^n = \text{int}\,S \cup \text{ext}\,S \cup \text{bd}\,S$.

Proof

(a) and (b) are obvious from definitions.
For parts (c) and (d), we notice that for a point $\mathbf{x} \in \mathbb{R}^n$, exactly one of the following three statements holds.

(i) There exists $r > 0$ such that $B(\mathbf{x}, r) \subset S$.

(ii) There exists $r > 0$ such that $B(\mathbf{x}, r) \subset \mathbb{R}^n \setminus S$.

(iii) For every $r > 0$, $B(\mathbf{x}, r)$ intersects both S and $\mathbb{R}^n \setminus S$.

Thus, $\text{int}\,S$, $\text{ext}\,S$ and $\text{bd}\,S$ are mutually disjoint sets, and their union is $\mathbb{R}^n$.

Example 1.30

Find the exterior and boundary of each of the following subsets of $\mathbb{R}$.

(a) $A = (a, b)$

(b) $B = (a, b]$

(c) $C = [a, b]$

(d) $\mathbb{Q}$

Solution

We have seen in Example 1.28 that

$$\text{int}\,A = \text{int}\,B = \text{int}\,C = (a, b).$$

For any $r > 0$, the ball $B(a, r) = (a - r, a + r)$ contains a point less than a, and a point larger than a. Hence, a is a boundary point of the sets A, B and C. Similarly, b is a boundary point of the sets A, B and C.
For every point x which satisfies $x < a$, let $r = a - x$. Then $r > 0$. Since $x + r = a$, the ball $B(x, r) = (x - r, x + r)$ is contained in $(-\infty, a)$. Hence, x is an exterior point of the sets A, B and C. Similarly every point x such that $x > b$ is an exterior point of the sets A, B and C.
Since the interior, exterior and boundary of a set in $\mathbb{R}$ are three mutually disjoint sets whose union is $\mathbb{R}$, we conclude that

$$\operatorname{bd} A = \operatorname{bd} B = \operatorname{bd} C = \{a, b\},$$
$$\operatorname{ext} A = \operatorname{ext} B = \operatorname{ext} C = (-\infty, a) \cup (b, \infty).$$

For every $x \in \mathbb{R}$ and every $r > 0$, the ball $B(x, r) = (x - r, x + r)$ contains a point in $\mathbb{Q}$ and a point not in $\mathbb{Q}$. Therefore, x is a boundary point of $\mathbb{Q}$. This shows that $\operatorname{bd} \mathbb{Q} = \mathbb{R}$, and thus, $\operatorname{ext} \mathbb{Q} = \emptyset$.

Example 1.31

Let $A = B(\mathbf{x}_0, r)$, where $\mathbf{x}_0$ is a point in $\mathbb{R}^n$, and r is a positive number. Find the interior, exterior and boundary of A.

Solution

We have shown that A is open. Hence, $\operatorname{int} A = A$. Let

$$U = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x} - \mathbf{x}_0\| > r\}, \qquad C = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x} - \mathbf{x}_0\| = r\}.$$

Notice that A, U and C are mutually disjoint sets whose union is $\mathbb{R}^n$.
If $\mathbf{x}$ is in U, $d = \|\mathbf{x} - \mathbf{x}_0\| > r$. Let $r' = d - r$. Then $r' > 0$. If $\mathbf{y} \in B(\mathbf{x}, r')$, then $\|\mathbf{y} - \mathbf{x}\| < r'$. It follows that

$$\|\mathbf{y} - \mathbf{x}_0\| \geq \|\mathbf{x} - \mathbf{x}_0\| - \|\mathbf{y} - \mathbf{x}\| > d - r' = r.$$

This proves that $\mathbf{y} \in U$. Hence, $B_d(\mathbf{x}, r') \subset U \subset \mathbb{R}^n \setminus A$, which shows that $\mathbf{x}$ is an exterior point of A. Thus, $U \subset \operatorname{ext} A$.

Now if $\mathbf{x} \in C$, $\|\mathbf{x} - \mathbf{x}_0\| = r$. For every $r' > 0$, let $a = \frac{1}{2}\min\{r'/r, 1\}$. Then $a \leq \frac{1}{2}$ and $a \leq \frac{r'}{2r}$. Consider the point

$$\mathbf{v} = \mathbf{x} - a(\mathbf{x} - \mathbf{x}_0).$$

Notice that

$$\|\mathbf{v} - \mathbf{x}\| = ar \leq \frac{r'}{2} < r'.$$

Thus, $\mathbf{v}$ is in $B(\mathbf{x}, r')$. On the other hand,

$$\|\mathbf{v} - \mathbf{x}_0\| = (1 - a)r < r.$$

Thus, $\mathbf{v}$ is in A. This shows that $B(\mathbf{x}, r')$ intersects A. Since $\mathbf{x}$ is in $B(\mathbf{x}, r')$ but not in A, we find that $B(\mathbf{x}, r')$ intersects $\mathbb{R}^n \setminus A$. Hence, $\mathbf{x}$ is a boundary point of A. This shows that $C \subset \operatorname{bd} A$.
Since $\operatorname{int} A$, $\operatorname{ext} A$ and $\operatorname{bd} A$ are mutually disjoint sets, we conclude that $\operatorname{int} A = A$, $\operatorname{ext} A = U$ and $\operatorname{bd} A = C$.

Now we introduce the closure of a set.

Definition 1.32 Closure

Let S be a subset of $\mathbb{R}^n$. The closure of S, denoted by $\overline{S}$, is defined as

$$\overline{S} = \operatorname{int} S \cup \operatorname{bd} S.$$

Example 1.32

Example 1.31 shows that the closure of the open ball $B(\mathbf{x}_0, r)$ is the closed ball $CB(\mathbf{x}_0, r)$.

Example 1.33

Consider the sets $A = (a, b)$, $B = (a, b]$ and $C = [a, b]$ in Example 1.28 and Example 1.30. We have shown that $\operatorname{int} A = \operatorname{int} B = \operatorname{int} C = (a, b)$, and $\operatorname{bd} A = \operatorname{bd} B = \operatorname{bd} C = \{a, b\}$. Therefore, $\overline{A} = \overline{B} = \overline{C} = [a, b]$.

Since $\mathbb{R}^n$ is a disjoint union of int S, bd S and ext S, we obtain the following immediately from the definition.

Theorem 1.31

Let S be a subset of $\mathbb{R}^n$. Then $\overline{S}$ and ext S are complement of each other in $\mathbb{R}^n$.

The following theorem gives a characterization of the closure of a set.

Theorem 1.32

Let S be a subset of $\mathbb{R}^n$, and let $\mathbf{x}$ be a point in $\mathbb{R}^n$. The following statements are equivalent.

(a) $\mathbf{x} \in \overline{S}$.

(b) For every $r > 0$, $B(\mathbf{x}, r)$ intersects S.

(c) There is a sequence $\{\mathbf{x}_k\}$ in S that converges to $\mathbf{x}$.

Proof

If $\mathbf{x}$ is in $\overline{S}$, $\mathbf{x}$ is not in int $(\mathbb{R}^n \setminus S)$. Thus, for every $r > 0$, $B(\mathbf{x}, r)$ is not contained in $\mathbb{R}^n \setminus S$. Then it must intersect S. This proves (a) implies (b).
If (b) holds, for every $k \in \mathbb{Z}^+$, take $r = 1/k$. The ball $B(\mathbf{x}, 1/k)$ intersects S at some point $\mathbf{x}_k$. This gives a sequence $\{\mathbf{x}_k\}$ satisfying

$$\|\mathbf{x}_k - \mathbf{x}\| < \frac{1}{k}.$$

Thus, $\{\mathbf{x}_k\}$ is a sequences in S that converges to $\mathbf{x}$. This proves (b) implies (c).
If (c) holds, for every $r > 0$, there is a positive integer K such that for all $k \geq K$, $\|\mathbf{x}_k - \mathbf{x}\| < r$, and thus $\mathbf{x}_k \in B(\mathbf{x}, r)$. This shows that $B(\mathbf{x}, r)$ is not contained in $\mathbb{R}^n \setminus S$. Hence, $\mathbf{x} \notin$ ext S, and thus we must have $\mathbf{x} \in \overline{S}$. This proves (c) implies (a).

The following theorem gives further properties of the closure of a set.

Theorem 1.33

Let S be a subset of $\mathbb{R}^n$.

1. $\overline{S}$ is a closed set that contains S.
2. S is closed if and only if $S = \overline{S}$.
3. If C is a closed subset of $\mathbb{R}^n$ and $S \subset C$, then $\overline{S} \subset C$.

These imply that $\overline{S}$ is the smallest closed set that contains S.

Proof

These statements are counterparts of the statements in Theorem 1.29.
Since $\operatorname{ext} S = \operatorname{int}(\mathbb{R}^n \setminus S)$, and the interior of a set is open, $\operatorname{ext} S$ is open. Since $\overline{S} = \mathbb{R}^n \setminus \operatorname{ext} S$, $\overline{S}$ is a closed set. Since $\operatorname{ext} S \subset \mathbb{R}^n \setminus S$, we find that

$$\overline{S} = \mathbb{R}^n \setminus \operatorname{ext} S \supset S.$$

If $S = \overline{S}$, then S must be closed since $\overline{S}$ is closed. Conversely, if S is closed, $\mathbb{R}^n \setminus S$ is open, and so $\operatorname{ext} S = \operatorname{int}(\mathbb{R}^n \setminus S) = \mathbb{R}^n \setminus S$. It follows that $\overline{S} = \mathbb{R}^n \setminus \operatorname{ext} S = S$.
If C is a closed set that contains S, then $\mathbb{R}^n \setminus C$ is an open set that is contained in $\mathbb{R}^n \setminus S$. Thus, $\mathbb{R}^n \setminus C \subset \operatorname{int}(\mathbb{R}^n \setminus S) = \operatorname{ext} S$. This shows that $C \supset \mathbb{R}^n \setminus \operatorname{ext} S = \overline{S}$.

Corollary 1.34

If S be a subset of $\mathbb{R}^n$, $\overline{S} = S \cup \operatorname{bd} S$.

Proof

Since $\operatorname{int} S \subset S$, $\overline{S} = \operatorname{int} S \cup \operatorname{bd} S \subset S \cup \operatorname{bd} S$. Since S and $\operatorname{bd} S$ are both subsets of $\overline{S}$, $S \cup \operatorname{bd} S \subset \overline{S}$. This proves that $\overline{S} = S \cup \operatorname{bd} S$.

Example 1.34

Let U be the open rectangle $U = \prod_{i=1}^{n}(a_i, b_i)$ in $\mathbb{R}^n$. Show that the closure of U is the closed rectangle $R = \prod_{i=1}^{n}[a_i, b_i]$.

Solution

Since R is a closed set that contains U, $\overline{U} \subset R$.
If $\mathbf{x} = (x_1, \ldots, x_n)$ is a point in R, then $x_i \in [a_i, b_i]$ for each $1 \leq i \leq n$. Since $[a_i, b_i]$ is the closure of (a_i, b_i) in $\mathbb{R}$, there is a sequence $\{x_{i,k}\}_{k=1}^{\infty}$ in (a_i, b_i) that converges to x_i. For $k \in \mathbb{Z}^+$, let

$$\mathbf{x}_k = (x_{1,k}, \ldots, x_{n,k}).$$

Then $\{\mathbf{x}_k\}$ is a sequence in U that converges to $\mathbf{x}$. This shows that $\mathbf{x} \in \overline{U}$, and thus completes the proof that $\overline{U} = R$.

The proof of the following theorem shows the usefulness of the characterization of int S as the largest open set that is contained in S, and $\overline{S}$ is the smallest closed set that contains S.

Theorem 1.35

If A and B are subsets of $\mathbb{R}^n$ such that $A \subset B$, then

(a) int $A \subset$ int B; and

(b) $\overline{A} \subset \overline{B}$.

Proof

Since int A is an open set that is contained in A, it is an open set that is contained in B. By the fourth statement in Theorem 1.29, int $A \subset$ int B.
Since $\overline{B}$ is a closed set that contains B, it is a closed set that contains A. By the third statement in Theorem 1.33, $\overline{A} \subset \overline{B}$.

Notice that as subsets of $\mathbb{R}$, $(a, b) \subset (a, b] \subset [a, b]$. We have shown that

$\overline{(a,b)} = \overline{(a,b]} = \overline{[a,b]}$. In general, we have the following.

Theorem 1.36

If A and B are subsets of $\mathbb{R}^n$ such that $A \subset B \subset \overline{A}$, then $\overline{A} = \overline{B}$.

Proof

By Theorem 1.35, $A \subset B$ implies that $\overline{A} \subset \overline{B}$, while $B \subset \overline{A}$ implies that $\overline{B}$ is contained in $\overline{\overline{A}} = \overline{A}$. Thus, we have

$$\overline{A} \subset \overline{B} \subset \overline{A},$$

which proves that $\overline{B} = \overline{A}$.

Example 1.35

In general, if S is a subset of $\mathbb{R}^n$, it is not necessary true that $\operatorname{int} S = \operatorname{int} \overline{S}$, even when S is an open set. For example, take $S = (-1,0) \cup (0,1)$ in $\mathbb{R}$. Then S is an open set and $\overline{S} = [-1,1]$. Notice that $\operatorname{int} S = S = (-1,0) \cup (0,1)$, but $\operatorname{int} \overline{S} = (-1,1)$.

Exercises 1.4

Question 1

Let S be a subset of $\mathbb{R}^n$. Show that $\operatorname{bd} S$ is a closed set.

Question 2

Let A be the subset of $\mathbb{R}^2$ given by

$$A = \{(x, y) \,|\, x < 0, y \geq 0\}\,.$$

Find the interior, exterior, boundary and closure of A.

Question 3

Let $\mathbf{x}_0$ be a point in $\mathbb{R}^n$, and let r be a positive number. Consider the subset of $\mathbb{R}^n$ given by

$$A = \{\mathbf{x} \in \mathbb{R}^n \,|\, 0 < \|\mathbf{x} - \mathbf{x}_0\| \leq r\}\,.$$

Find the interior, exterior, boundary and closure of A.

Question 4

Let A be the subset of $\mathbb{R}^2$ given by

$$A = \{(x, y) \,|\, 1 \leq x < 3, -2 < y \leq 5\} \cup \{(0, 0), (2, -3)\}.$$

Find the interior, exterior, boundary and closure of A.

Question 5

Let S be a subset of $\mathbb{R}^n$. Show that

$$\operatorname{bd} S = \overline{S} \cap \overline{\mathbb{R}^n \setminus S}.$$

Question 6

Let S be a subset of $\mathbb{R}^n$. Show that $\operatorname{bd}\overline{S} \subset \operatorname{bd} S$. Give an example where $\operatorname{bd}\overline{S} \neq \operatorname{bd} S$.

Question 7

Let S be a subset of $\mathbb{R}^n$.

(a) Show that S is open if and only if S does not contain any of its boundary points.

(b) Show that S is closed if and only if S contains all its boundary points.

Question 8

Let S be a subset of $\mathbb{R}^n$, and let $\mathbf{x}$ be a point in $\mathbb{R}^n$.

(a) Show that $\mathbf{x}$ is an interior point of S if and only if there is a neighbourhood of $\mathbf{x}$ that is contained in S.

(b) Show that $\mathbf{x} \in \overline{S}$ if and only if every neighbourhood of $\mathbf{x}$ intersects S.

(c) Show that $\mathbf{x}$ is a boundary point of S if and only if every neighbourhood of $\mathbf{x}$ contains a point in S and a point not in S.

Question 9

Let S be a subset of $\mathbb{R}^n$, and let $\mathbf{x} = (x_1, \ldots, x_n)$ be a point in the interior of S.

(a) Show that there is an $r_1 > 0$ such that $CB(\mathbf{x}, r_1) \subset S$.

(b) Show that there is an $r_2 > 0$ such that $\displaystyle\prod_{i=1}^{n}(x_i - r_2, x_i + r_2) \subset S$.

(c) Show that there is an $r_3 > 0$ such that $\displaystyle\prod_{i=1}^{n}[x_i - r_3, x_i + r_3] \subset S$.

1.5 Limit Points and Isolated Points

In this section, we generalize the concepts of limit points and isolated points to subsets of $\mathbb{R}^n$.

Definition 1.33 Limit Points

Let S be a subset of $\mathbb{R}^n$. A point $\mathbf{x}$ in $\mathbb{R}^n$ is a limit point of S provided that there is a sequence $\{\mathbf{x}_k\}$ in $S \setminus \{\mathbf{x}\}$ that converges to $\mathbf{x}$. The set of limit points of S is denoted by S'.

By Theorem 1.32, we obtain the following immediately.

Theorem 1.37

Let S be a subset of $\mathbb{R}^n$, and let $\mathbf{x}$ be a point in $\mathbb{R}^n$. The following are equivalent.

(a) $\mathbf{x}$ is a limit point of S.

(b) $\mathbf{x}$ is in $\overline{S \setminus \{\mathbf{x}\}}$.

(c) For every $r > 0$, $B(\mathbf{x}, r)$ intersects S at a point other than $\mathbf{x}$.

Corollary 1.38

If S is a subset of $\mathbb{R}^n$, then $S' \subset \overline{S}$.

Proof

If $\mathbf{x} \in S'$, $\mathbf{x} \in \overline{S \setminus \{\mathbf{x}\}}$. Since $S \setminus \{\mathbf{x}\} \subset S$, we have $\overline{S \setminus \{\mathbf{x}\}} \subset \overline{S}$. Therefore, $\mathbf{x} \in \overline{S}$.

The following theorem says that the closure of a set is the union of the set with all its limit points.

Theorem 1.39

If S is a subset of $\mathbb{R}^n$, then $\overline{S} = S \cup S'$.

Proof

By Corollary 1.38, $S' \subset \overline{S}$. Since we also have $S \subset \overline{S}$, we find that $S \cup S' \subset \overline{S}$.
Conversely, if $\mathbf{x} \in \overline{S}$, then by Theorem 1.32, there is a sequence $\{\mathbf{x}_k\}$ in S that converges to $\mathbf{x}$. If $\mathbf{x}$ is not in S, then the sequence $\{\mathbf{x}_k\}$ is in $S \setminus \{\mathbf{x}\}$. In this case, $\mathbf{x}$ is a limit point of S. This shows that $\overline{S} \setminus S \subset S'$, and hence, $\overline{S} \subset S \cup S'$.

In the proof above, we have shown the following.

Corollary 1.40

Let S be a subset of $\mathbb{R}^n$. Every point in $\overline{S}$ that is not in S is a limit point of S. Namely,

$$\overline{S} \setminus S \subset S'.$$

Now we introduce the definition of isolated points.

Definition 1.34 Isolated Points

Let S be a subset of $\mathbb{R}^n$. A point $\mathbf{x}$ in $\mathbb{R}^n$ is an isolated point of S if

(a) $\mathbf{x}$ is in S;

(b) $\mathbf{x}$ is not a limit point of S.

Remark 1.1

By definition, a point $\mathbf{x}$ in S is either an isolated point of S or a limit point of S.

Theorem 1.37 gives the following immediately.

Theorem 1.41

Let S be a subset of $\mathbb{R}^n$ and let $\mathbf{x}$ be a point in S. Then $\mathbf{x}$ is an isolated point of S if and only if there is an $r > 0$ such that the ball $B(\mathbf{x}, r)$ does not contain other points of S except the point $\mathbf{x}$.

Example 1.36

Find the set of limit points and isolated points of the set $A = \mathbb{Z}^2$ as a subset of $\mathbb{R}^2$.

Solution

If $\{\mathbf{x}_k\}$ is a sequence in A that converges to a point $\mathbf{x}$, then there is a positive integer K such that for all $l \geq k \geq K$,

$$\|\mathbf{x}_l - \mathbf{x}_k\| < 1.$$

This implies that $\mathbf{x}_k = \mathbf{x}_K$ for all $k \geq K$. Hence, $\mathbf{x} = \mathbf{x}_K \in A$. This shows that A is closed. Hence, $\overline{A} = A$. Therefore, $A' \subset A$.
For every $\mathbf{x} = (k, l) \in \mathbb{Z}^2$, $B(\mathbf{x}, 1)$ intersects A only at the point $\mathbf{x}$ itself. Hence, $\mathbf{x}$ is an isolated point of A. This shows that every point of A is an isolated point. Since $A' \subset A$, we must have $A' = \emptyset$.

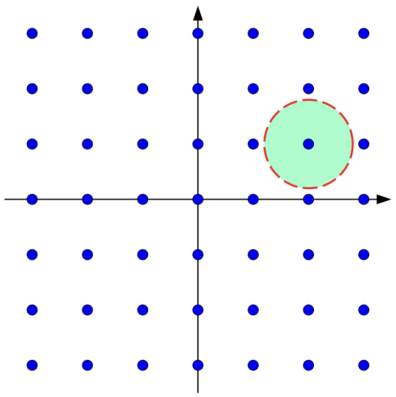


Figure 1.17: The set $\mathbb{Z}^2$ does not have limit points.

Let us prove the following useful fact.

Theorem 1.42

If S is a subset of $\mathbb{R}^n$, every interior point of S is a limit point of S.

Proof

If $\mathbf{x}$ is an interior point of S, there exists $r_0 > 0$ such that $B(\mathbf{x}, r_0) \subset S$. Given $r > 0$, let $r' = \frac{1}{2}\min\{r, r_0\}$. Then $r' > 0$. Since $r' < r$ and $r' < r_0$, the point

$$\mathbf{x}' = \mathbf{x} + r'\mathbf{e}_1$$

is in $B(\mathbf{x}, r)$ and S. Obviously, $\mathbf{x}' \neq \mathbf{x}$. Therefore, for every $r > 0$, $B(\mathbf{x}, r)$ intersects S at a point other than $\mathbf{x}$. This proves that $\mathbf{x}$ is a limit point of S.

Since $S \subset \operatorname{int} S \cup \operatorname{bd} S$, and $\operatorname{int} S$ and $\operatorname{bd} S$ are disjoint, we deduce the following.

Corollary 1.43

Let S be a subset of $\mathbb{R}^n$. An isolated point of S must be a boundary point.

Since every point in an open set S is an interior point of S, we obtain the following.

Corollary 1.44

If S is an open subset of $\mathbb{R}^n$, every point of S is a limit point. Namely, $S \subset S'$.

Example 1.37

If I is an interval of the form (a, b), $(a, b]$, $[a, b)$ or $[a, b]$ in $\mathbb{R}$, then $\operatorname{bd} I = \{a, b\}$. It is easy to check that a and b are not isolated points of I. Hence, I has no isolated points. Since $\overline{I} = I \cup I'$ and $I \subset I'$, we find that $I' = \overline{I} = [a, b]$.

In fact, we can prove a general theorem.

Theorem 1.45

Let A and B be subsets of $\mathbb{R}^n$ such that A is open and $A \subset B \subset \overline{A}$. Then $B' = \overline{A}$. In particular, the set of limit points of A is $\overline{A}$.

Proof

By Theorem 1.36, $\overline{A} = \overline{B}$. Since A is open, $A \subset A'$. Since $\overline{A} = A \cup A'$, we find that $\overline{A} = A'$.
In the exercises, one is asked to show that $A \subset B$ implies $A' \subset B'$. Therefore, $\overline{A} = A' \subset B' \subset \overline{B}$. Since $\overline{A} = \overline{B}$, we must have $B' = \overline{B} = \overline{A}$.

Example 1.38

Let A be the subset of $\mathbb{R}^2$ given by

$$A = [-1, 1] \times (-2, 2] = \{(x, y) \mid -1 \leq x \leq 1, -2 < y \leq 2\}.$$

Since $U = (-1, 1) \times (-2, 2)$ is open, $\overline{U} = [-1, 1] \times [-2, 2]$, and $U \subset A \subset \overline{U}$, the set of limit points of A is $\overline{U} = [-1, 1] \times [-2, 2]$.

Exercises 1.5

Question 1

Let A and B be subsets of $\mathbb{R}^n$ such that $A \subset B$. Show that $A' \subset B'$.

Question 2

Let $\mathbf{x}_0$ be a point in $\mathbb{R}^n$ and let r be a positive number. Find the set of limit points of the open ball $B(\mathbf{x}_0, r)$.

Question 3

Let A be the subset of $\mathbb{R}^2$ given by

$$A = \{(x, y) \,|\, x < 0, y \geq 0\}\,.$$

Find the set of limit points of A.

Question 4

Let $\mathbf{x}_0$ be a point in $\mathbb{R}^n$, and let r is a positive number. Consider the subset of $\mathbb{R}^n$ given by

$$A = \{\mathbf{x} \in \mathbb{R}^n \,|\, 0 < \|\mathbf{x} - \mathbf{x}_0\| \leq r\}\,.$$

(a) Find the set of limit points of A.

(b) Find the set of isolated points of the set $S = \mathbb{R}^n \setminus A$.

Question 5

Let A be the subset of $\mathbb{R}^2$ given by

$$A = \{(x, y) \,|\, 1 \leq x < 3, -2 < y \leq 5\} \cup \{(0, 0), (2, -3)\}.$$

Determine the set of isolated points and the set of limit points of A.

Question 6

Let $A = \mathbb{Q}^2$ as a subset of $\mathbb{R}^2$.

(a) Find the interior, exterior, boundary and closure of A.

(b) Determine the set of isolated points and the set of limit points of A.

Question 7

Let S be a subset of $\mathbb{R}^n$. Show that S is closed if and only if it contains all its limit points.

Question 8

Let S be a subset of $\mathbb{R}^n$, and let $\mathbf{x}$ be a point in $\mathbb{R}^n$. Show that $\mathbf{x}$ is a limit point of S if and only if every neighbourhood of $\mathbf{x}$ intersects S at a point other than itself.

Question 9

Let $\mathbf{x}_1$, $\mathbf{x}_2$, . . ., $\mathbf{x}_k$ be points in $\mathbb{R}^n$ and let $A = \mathbb{R}^n \setminus \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$. Find the set of limit points of A.

Chapter 2

Limits of Multivariable Functions and Continuity

We are interested in functions $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ that are defined on subsets $\mathfrak{D}$ of $\mathbb{R}^n$, taking values in $\mathbb{R}^m$. When $n \geq 2$, these are called multivariable functions. When $m \geq 2$, they are called vector-valued functions. When $m = 1$, we usually write the function as $f : \mathfrak{D} \to \mathbb{R}$.

2.1 Multivariable Functions

In this section, let us define some special classes of multivariable functions.

2.1.1 Polynomials and Rational Functions

A special class of functions is the set of polynomials in n variables.

Definition 2.1 Polynomials

Let $\mathbf{k} = (k_1, \ldots, k_n)$ be an n-tuple of nonnegative integers. Associated to this n-tuple $\mathbf{k}$, there is a monomial $p_{\mathbf{k}} : \mathbb{R}^n \to \mathbb{R}$ of degree $|\mathbf{k}| = k_1 + \cdots + k_n$ of the form $p_{\mathbf{k}}(\mathbf{x}) = x_1^{k_1} \cdots x_n^{k_n}$.
A polynomial in n variables is a function $p : \mathbb{R}^n \to \mathbb{R}$ that is a *finite* linear combination of monomials in n variables. It takes the form

$$p(\mathbf{x}) = \sum_{j=1}^{m} c_{\mathbf{k}_j} p_{\mathbf{k}_j}(\mathbf{x}),$$

where $\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_m$ are distinct n-tuples of nonnegative integers, and $c_{\mathbf{k}_1}, c_{\mathbf{k}_2}, \ldots, c_{\mathbf{k}_m}$ are nonzero real numbers. The degree of the polynomial $p(\mathbf{x})$ is $\max\{|\mathbf{k}_1|, |\mathbf{k}_2|, \ldots, |\mathbf{k}_m|\}$.

Example 2.1

The following are examples of polynomials in three variables.

(a) $p(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2$

(b) $p(x_1, x_2, x_3) = 4x_1^2x_2 - 3x_1x_3 + x_1x_2x_3$

Example 2.2

The function $f : \mathbb{R}^n \to \mathbb{R}$,

$$f(\mathbf{x}) = \|\mathbf{x}\| = \sqrt{x_1^2 + \cdots + x_n^2}$$

is not a polynomial.

When the domain of a function is not specified, we always assume that the domain is the largest set on which the function can be defined.

Definition 2.2 Rational Functions

A rational function $f : \mathfrak{D} \to \mathbb{R}$ is the quotient of two polynomials $p : \mathbb{R}^n \to \mathbb{R}$ and $q : \mathbb{R}^n \to \mathbb{R}$. Namely,

$$f(\mathbf{x}) = \frac{p(\mathbf{x})}{q(\mathbf{x})}.$$

Its domain $\mathfrak{D}$ is the set

$$\mathfrak{D} = \{\mathbf{x} \in \mathbb{R}^n \,|\, q(\mathbf{x}) \neq 0\}.$$

Example 2.3

The function

$$f(x_1, x_2) = \frac{x_1x_2 + 3x_1^2}{x_1 - x_2}$$

is a rational function defined on the set

$$\mathfrak{D} = \left\{(x_1, x_2) \in \mathbb{R}^2 \,|\, x_1 \neq x_2\right\}.$$

2.1.2 Component Functions of a Mapping

If the codomain $\mathbb{R}^m$ of the function $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ has dimension $m \geq 2$, we usually call the function a *mapping*. In this case, it would be good to consider the component functions.

For $1 \leq j \leq m$, the projection function $\pi_j : \mathbb{R}^m \to \mathbb{R}$ is the function

$$\pi_j(x_1, \ldots, x_m) = x_j.$$

Definition 2.3 Component Functions

Let $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ be a function defined on $\mathfrak{D} \subset \mathbb{R}^n$. For $1 \leq j \leq m$, the j^{th} component function of $\mathbf{F}$ is the function $F_j : \mathfrak{D} \to \mathbb{R}$ defined as

$$F_j = (\pi_j \circ \mathbf{F}) : \mathfrak{D} \to \mathbb{R}.$$

For each $\mathbf{x} \in \mathfrak{D}$,

$$\mathbf{F}(\mathbf{x}) = (F_1(\mathbf{x}), \ldots, F_m(\mathbf{x})).$$

Example 2.4

For the function $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^3$, $\mathbf{F}(\mathbf{x}) = -3\mathbf{x}$, the component functions are $F_1(x_1, x_2, x_3) = -3x_1$, $F_2(x_1, x_2, x_3) = -3x_2$, $F_3(x_1, x_2, x_3) = -3x_3$.

For convenience, we also define the notion of polynomial mappings.

Definition 2.4 Polynomial Mappings

We call a function $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$ a polynomial mapping if each of its components $F_j : \mathbb{R}^n \to \mathbb{R}$, $1 \leq j \leq m$, is a polynomial function. The degree of the polynomial mapping $\mathbf{F}$ is the maximum of the degrees of the polynomials $F_1, F_2, \ldots, F_m$.

Example 2.5

The mapping $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^2$,

$$\mathbf{F}(x, y, z) = (x^2y + 3xz, 8yz^3 - 7x)$$

is a polynomial mapping of degree 4.

2.1.3 Invertible Mappings

The invertibility of a function $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is defined in the following way.

Definition 2.5 Inverse Functions

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$, and let $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ be a function defined on $\mathfrak{D}$. We say that $\mathbf{F}$ is invertible if $\mathbf{F}$ is one-to-one. In this case, the inverse function $\mathbf{F}^{-1} : \mathbf{F}(\mathfrak{D}) \to \mathfrak{D}$ is defined so that for each $\mathbf{y} \in \mathbf{F}(\mathfrak{D})$,

$$\mathbf{F}^{-1}(\mathbf{y}) = \mathbf{x} \qquad \text{if and only if} \qquad \mathbf{F}(\mathbf{x}) = \mathbf{y}.$$

Example 2.6

Let $\mathfrak{D} = \{(x, y) \,|\, x > 0, y > 0\}$ and let $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^2$ be the function defined as

$$\mathbf{F}(x, y) = (x - y, x + y).$$

Show that $\mathbf{F}$ is invertible and find its inverse.

Solution

Let $u = x - y$ and $v = x + y$. Then

$$x = \frac{u + v}{2}, \qquad y = \frac{v - u}{2}.$$

This shows that for any $(u, v) \in \mathbb{R}^2$, there is at most one pair of (x, y) such that $\mathbf{F}(x, y) = (u, v)$. Thus, $\mathbf{F}$ is one-to-one, and hence, it is invertible. Observe that

$$\mathbf{F}(\mathfrak{D}) = \{(u, v) \,|\, v > 0, -v < u < v.\}\,.$$

The inverse mapping is given by $\mathbf{F}^{-1} : \mathbf{F}(\mathfrak{D}) \to \mathbb{R}^2$,

$$\mathbf{F}^{-1}(u, v) = \left(\frac{u + v}{2}, \frac{v - u}{2}\right).$$

2.1.4 Linear Transformations

Another special class of functions consists of linear transformations. A function $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation if for any $\mathbf{x}_1, \ldots, \mathbf{x}_k$ in $\mathbb{R}^n$, and for any $c_1, \ldots, c_k$ in $\mathbb{R}$,

$$\mathbf{T}(c_1\mathbf{x}_1 + \cdots + c_k\mathbf{x}_k) = c_1\mathbf{T}(\mathbf{x}_1) + \cdots + c_k\mathbf{T}(\mathbf{x}_k).$$

Linear transformations are closely related to matrices.

An $m \times n$ matrix A is an array with m rows and n columns of real numbers. It has the form

$$A = [a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

If $A = [a_{ij}]$ and $B = [b_{ij}]$ are $m \times n$ matrices, α and β are real numbers, $\alpha A + \beta B$ is defined to be the $m \times n$ matrix $C = \alpha A + \beta B = [c_{ij}]$ with

$$c_{ij} = \alpha a_{ij} + \beta b_{ij}.$$

If $A = [a_{il}]$ is a $m \times k$ matrix, $B = [b_{lj}]$ is a $k \times n$ matrix, the product AB is defined to be the $m \times n$ matrix $C = AB = [c_{ij}]$, where

$$c_{ij} = \sum_{l=1}^{k} a_{il} b_{lj}.$$

It is easy to verify that matrix multiplications are associative.

Given $\mathbf{x} = (x_1, \ldots, x_n)$ in $\mathbb{R}^n$, we identify it with the column vector

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix},$$

which is an $n \times 1$ matrix. If A is an $m \times n$ matrix, and $\mathbf{x}$ is a vector in $\mathbb{R}^n$, then $\mathbf{y} = A\mathbf{x}$ is the vector in $\mathbb{R}^m$ given by

$$\mathbf{y} = A\mathbf{x} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{bmatrix}.$$

The following is a standard result in linear algebra.

Theorem 2.1

A function $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation if and only if there exists an $m \times n$ matrix $A = [a_{ij}]$ such that

$$\mathbf{T}(\mathbf{x}) = A\mathbf{x}.$$

In this case, A is called the matrix associated to the linear transformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^m$.

Sketch of Proof

It is easy to verify that the mapping $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{T}(\mathbf{x}) = A\mathbf{x}$ is a linear transformation if A is an $m \times n$ matrix.
Conversely, if $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation, then for any $\mathbf{x} \in \mathbb{R}^n$,

$$\mathbf{T}(\mathbf{x}) = \mathbf{T}(x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_n\mathbf{e}_n) = x_1\mathbf{T}(\mathbf{e}_1) + x_2\mathbf{T}(\mathbf{e}_2) + \cdots + x_n\mathbf{T}(\mathbf{e}_n).$$

Define the vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ in $\mathbb{R}^m$ by

$$\mathbf{a}_1 = \mathbf{T}(\mathbf{e}_1),\ \mathbf{a}_2 = \mathbf{T}(\mathbf{e}_2),\ \ldots,\ \mathbf{a}_n = \mathbf{T}(\mathbf{e}_n).$$

Let A be the $m \times n$ matrix with column vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$. Namely,

$$A = \left[\mathbf{a}_1 \,\middle|\, \mathbf{a}_2 \,\middle|\, \cdots \,\middle|\, \mathbf{a}_n\right].$$

Then we have $\mathbf{T}(\mathbf{x}) = A\mathbf{x}$.

Example 2.7

Let $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$ be the function defined as

$$\mathbf{F}(x, y) = (x - y, x + y).$$

Then $\mathbf{F}$ is a linear transformation with matrix $A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$.

For the linear transformation $T : \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{T}(\mathbf{x}) = A\mathbf{x}$, the component functions are

$$\begin{aligned} T_1(\mathbf{x}) &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n, \\ T_2(\mathbf{x}) &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n, \\ &\vdots \\ T_m(\mathbf{x}) &= a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n. \end{aligned}$$

Each of them is a polynomial of degree at most one. Thus, a linear transformation is a polynomial mapping of degree at most one. It is easy to deduce the following.

Corollary 2.2

A mapping $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation if and only if each component function is a linear transformation.

The followings are some standard results about linear transformations.

Theorem 2.3

If $\mathbf{S} : \mathbb{R}^n \to \mathbb{R}^m$ and $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^m$ are linear transformations with matrices A and B respectively, then for any real numbers α and β, $\alpha\mathbf{S} + \beta\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation with matrix $\alpha A + \beta B$.

Theorem 2.4

If $\mathbf{S} : \mathbb{R}^n \to \mathbb{R}^m$ and $\mathbf{T} : \mathbb{R}^m \to \mathbb{R}^k$ are linear transformations with matrices A and B, then $\mathbf{T} \circ \mathbf{S} : \mathbb{R}^n \to \mathbb{R}^k$ is a linear transformation with matrix BA.

Sketch of Proof

This follows from

$$(\mathbf{T} \circ \mathbf{S})(\mathbf{x}) = \mathbf{T}(\mathbf{S}(\mathbf{x})) = B(A\mathbf{x}) = (BA)\mathbf{x}.$$

In the particular case when $m = n$, we have the following.

Theorem 2.5

Let $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$ be a linear transformation represented by the matrix A. The following are equivalent.

(a) The mapping $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$ is one-to-one.

(b) The mapping $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$ is onto.

(c) The matrix A is invertible.

(d) $\det A \neq 0$.

In other words, if the linear transformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$ is one-to-one *or* onto, then it is bijective. In this case, the linear transformation is invertible, and we can define the inverse function $\mathbf{T}^{-1} : \mathbb{R}^n \to \mathbb{R}^n$.

Theorem 2.6

Let $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$ be an invertible linear transformation represented by the matrix A. Then the inverse mapping $\mathbf{T}^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ is also a linear transformation and

$$\mathbf{T}^{-1}(\mathbf{x}) = A^{-1}\mathbf{x}.$$

Example 2.8

Let $\mathbf{T} : \mathbb{R}^2 \to \mathbb{R}^2$ be the linear transformation

$$\mathbf{T}(x, y) = (x - y, x + y).$$

The matrix associated with $\mathbf{T}$ is $A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$. Since $\det A = 2 \neq 0$, $\mathbf{T}$ is invertible. Since $A^{-1} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$, we have

$$\mathbf{T}^{-1}(x, y) = \left(\frac{x + y}{2}, \frac{-x + y}{2}\right).$$

2.1.5 Quadratic Forms

Given an $m \times n$ matrix $A = [a_{ij}]$, its *transpose* is the $n \times m$ matrix $A^T = [b_{ij}]$, where

$$b_{ij} = a_{ji} \qquad \text{for all } 1 \leq i \leq n, 1 \leq j \leq m.$$

An $n \times n$ matrix A is **symmetric** if

$$A = A^T.$$

An $n \times n$ matrix P is **orthogonal** if

$$P^T P = PP^T = I.$$

If the column vectors of P are $\mathbf{v}_1$, $\mathbf{v}_2$, ..., $\mathbf{v}_n$, so that

$$P = \left[\mathbf{v}_1 \;\middle|\; \mathbf{v}_2 \;\middle|\; \cdots \;\middle|\; \mathbf{v}_n\right], \tag{2.1}$$

then P is orthogonal if and only if $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is an orthonormal set of vectors in $\mathbb{R}^n$.

If A is an $n \times n$ symmetric matrix, its characteristic polynomial

$$p(\lambda) = \det(\lambda I_n - A)$$

is a monic polynomial of degree n with n real roots $\lambda_1, \lambda_2, \ldots, \lambda_n$, counting with multiplicities. These roots are called the **eigenvalues** of A. There is an orthonormal set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ in $\mathbb{R}^n$ such that

$$A\mathbf{v}_i = \lambda_i \mathbf{v}_i \qquad \text{for all } 1 \leq i \leq n. \tag{2.2}$$

Let D be the diagonal matrix

$$D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}, \tag{2.3}$$

and let P be the orthogonal matrix (2.1). Then (2.2) is equivalent to $AP = PD$, or equivalently,

$$A = PDP^T = PDP^{-1}.$$

This is known as the orthogonal diagonalization of the real symmetric matrix A.

A quadratic form in $\mathbb{R}^n$ is a polynomial function $Q : \mathbb{R}^n \to \mathbb{R}$ of the form

$$Q(\mathbf{x}) = \sum_{1 \le i < j \le n} c_{ij} x_i x_j.$$

An $n \times n$ symmetric matrix $A = [a_{ij}]$ defines a quadratic form $Q_A : \mathbb{R}^n \to \mathbb{R}$ by

$$Q_A(\mathbf{x}) = \mathbf{x}^T A \mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j.$$

Example 2.9

The symmetric matrix $A = \begin{bmatrix} 1 & -2 \\ -2 & 5 \end{bmatrix}$ defines the quadratic form

$$Q_A(x, y) = x^2 - 4xy + 5y^2.$$

Conversely, given a quadratic form

$$Q(\mathbf{x}) = \sum_{1 \le i < j \le n} c_{ij} x_i x_j,$$

then $Q = Q_A$, where the entries of $A = [a_{ij}]$ are

$$a_{ij} = \begin{cases} c_{ii}, & \text{if } i = j, \\ c_{ij}/2, & \text{if } i < j, \\ c_{ji}/2, & \text{if } i > j. \end{cases}$$

Thus, there is a one-to-one correspondence between quadratic forms and symmetric matrices.

If $A = PDP^T$ is an orthogonal diagonalization of A, under the change of variables

$$\mathbf{y} = P^T \mathbf{x}, \quad \text{or equivalently,} \quad \mathbf{x} = P\mathbf{y}$$

we find that

$$Q_A = \mathbf{y}^T D \mathbf{y} = \lambda_1 y_1^2 + \cdots + \lambda_n y_n^2. \tag{2.4}$$

A consequence of (2.4) is the following.

Theorem 2.7

Let A be an $n \times n$ symmetric matrix, and let $Q_A(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ be the associated quadratic form. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of A. Assume that

$$\lambda_n \leq \cdots \leq \lambda_2 \leq \lambda_1.$$

Then for any $\mathbf{x} \in \mathbb{R}^n$,

$$\lambda_n \|\mathbf{x}\|^2 \leq Q_A(\mathbf{x}) \leq \lambda_1 \|\mathbf{x}\|^2.$$

Sketch of Proof

Given $\mathbf{x} \in \mathbb{R}^n$, let $\mathbf{y} = P^T\mathbf{x}$. Then

$$\|\mathbf{y}\|^2 = \mathbf{y}^T\mathbf{y} = \mathbf{x}^T PP^T\mathbf{x} = \mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|^2.$$

By (2.4) ,

$$Q_A(\mathbf{x}) = \lambda_1 y_1^2 + \cdots + \lambda_n y_n^2.$$

Since $\lambda_n \leq \cdots \leq \lambda_2 \leq \lambda_1$, we find that

$$\lambda_n(y_1^2 + \ldots + y_n^2) \leq Q_A(\mathbf{x}) \leq \lambda_1(y_1^2 + \ldots + y_n^2).$$

The assertion follows.

At the end of this section, let us recall the classification of quadratic forms.

Definiteness of Symmetric Matrices

Given an $n \times n$ symmetric matrix $A = [a_{ij}]$, let $Q_A : \mathbb{R}^n \to \mathbb{R}$,

$$Q_A(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j$$

be the associated quadratic form.

1. We say that the matrix A is positive definite, or the quadratic form Q_A is positive definite, if $Q_A(\mathbf{x}) > 0$ for all $\mathbf{x} \neq \mathbf{0}$ in $\mathbb{R}^n$.

2. We say that the matrix A is negative definite, or the quadratic form Q_A is negative definite, if $Q_A(\mathbf{x}) < 0$ for all $\mathbf{x} \neq \mathbf{0}$ in $\mathbb{R}^n$.

3. We say that the matrix A is indefinite, or the quadratic form Q_A is indefinite, if there exist $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$ such that $Q_A(\mathbf{u}) > 0$ and $Q_A(\mathbf{v}) < 0$.

4. We say that the matrix A is positive semi-definite, or the quadratic form Q_A is positive semi-definite, if $Q_A(\mathbf{x}) \geq 0$ for all $\mathbf{x}$ in $\mathbb{R}^n$.

5. We say that the matrix A is negative semi-definite, or the quadratic form Q_A is negative semi-definite, if $Q_A(\mathbf{x}) \leq 0$ for all $\mathbf{x}$ in $\mathbb{R}^n$.

Obviously, a symmetric matrix A is negative definite if and only if $-A$ is positive definite.

The following is a standard result in linear algebra, which can be deduced from (2.4).

Theorem 2.8

Let A be an $n \times n$ symmetric matrix, and let $Q_A(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ be the associated quadratic form. Let $\{\lambda_1, \ldots, \lambda_n\}$ be the set of eigenvalues of A, repeated with multiplicities.

(a) Q_A is positive definite if and only if $\lambda_i > 0$ for all $1 \leq i \leq n$.

(b) Q_A is negative definite if and only if $\lambda_i < 0$ for all $1 \leq i \leq n$.

(c) Q_A is indefinite if there exist i and j so that $\lambda_i > 0$ and $\lambda_j < 0$.

(d) Q_A is positive semi-definite if and only if $\lambda_i \geq 0$ for all $1 \leq i \leq n$.

(e) Q_A is negative semi-definite if and only if $\lambda_i \leq 0$ for all $1 \leq i \leq n$.

From Theorem 2.7 and Theorem 2.8, we obtain the following.

Corollary 2.9

Let $Q : \mathbb{R}^n \to \mathbb{R}$ be a quadratic form. If Q is positive definite, then there exists a positive constant c such that

$$Q(\mathbf{x}) \geq c\|\mathbf{x}\|^2 \qquad \text{for all } \mathbf{x} \in \mathbb{R}^n.$$

In fact, c can be any positive number that is less than or equal to the smallest eigenvalue of the symmetric matrix A associated to the quadratic form Q.

2.2 Limits of Functions

In this section, we study limits of multivariable functions.

Definition 2.6 Limits of Functions

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$ and let $\mathbf{x}_0$ be a limit point of $\mathfrak{D}$. Given a function $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$, we say that *the limit of* $\mathbf{F}(\mathbf{x})$ *as* $\mathbf{x}$ *approaches* $\mathbf{x}_0$ *is* $\mathbf{v}$, provided that whenever $\{\mathbf{x}_k\}$ is a sequence of points in $\mathfrak{D} \setminus \{\mathbf{x}_0\}$ that converges to $\mathbf{x}_0$, the sequence $\{\mathbf{F}(\mathbf{x}_k)\}$ of points in $\mathbb{R}^m$ converges to the point $\mathbf{v}$.

If the limit of $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ as $\mathbf{x}$ approaches $\mathbf{x}_0$ is $\mathbf{v}$, we write

$$\lim_{\mathbf{x} \to \mathbf{x}_0} \mathbf{F}(\mathbf{x}) = \mathbf{v}.$$

Example 2.10

For $1 \leq i \leq n$, let $\pi_i : \mathbb{R}^n \to \mathbb{R}$ be the projection function $\pi_i(x_1, \ldots, x_n) = x_i$. By the theorem on componentwise convergence of sequences, if $\{\mathbf{x}_k\}$ is a sequence in $\mathbb{R}^n \setminus \{\mathbf{x}_0\}$ that converges to the point $\mathbf{x}_0$, then

$$\lim_{k \to \infty} \pi_i(\mathbf{x}_k) = \pi_i(\mathbf{x}_0).$$

This means that

$$\lim_{\mathbf{x} \to \mathbf{x}_0} \pi_i(\mathbf{x}) = \pi_i(\mathbf{x}_0).$$

From the theorem on componentwise convergence of sequences, we also obtain the following immediately.

Proposition 2.10

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$ and let $\mathbf{x}_0$ be a limit point of $\mathfrak{D}$. Given a function $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$,

$$\lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{F}(\mathbf{x}) = \mathbf{v}$$

if and only if for each $1 \leq j \leq m$,

$$\lim_{\mathbf{x}\to\mathbf{x}_0} F_j(\mathbf{x}) = \pi_j(\mathbf{v}).$$

Example 2.11

Let $f : \mathbb{R}^n \to \mathbb{R}$ be the function defined as $f(\mathbf{x}) = \|\mathbf{x}\|$. If $\mathbf{x}_0$ is a point in $\mathbb{R}^n$, find $\lim_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x})$.

Solution

We have shown in Example 1.15 that If $\{\mathbf{x}_k\}$ is a sequence in $\mathbb{R}^n \setminus \{\mathbf{x}_0\}$ that converges to $\mathbf{x}_0$, then

$$\lim_{k\to\infty} \|\mathbf{x}_k\| = \|\mathbf{x}_0\|.$$

Therefore, $\lim_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x}) = \|\mathbf{x}_0\|$.

By the limit laws for sequences, we also have the followings.

Proposition 2.11

Let $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ and $\mathbf{G} : \mathfrak{D} \to \mathbb{R}^m$ be functions defined on $\mathfrak{D} \subset \mathbb{R}^n$. If $\mathbf{x}_0$ is a limit point of $\mathfrak{D}$ and

$$\lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{F}(\mathbf{x}) = \mathbf{u}, \qquad \lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{G}(\mathbf{x}) = \mathbf{v},$$

then for any real numbers α and β,

$$\lim_{\mathbf{x}\to\mathbf{x}_0} (\alpha\mathbf{F} + \beta\mathbf{G})(\mathbf{x}) = \alpha\mathbf{u} + \beta\mathbf{v}.$$

Proposition 2.12

Let $f : \mathfrak{D} \to \mathbb{R}$ and $g : \mathfrak{D} \to \mathbb{R}$ be functions defined on $\mathfrak{D} \subset \mathbb{R}^n$. If $\mathbf{x}_0$ is a limit point of $\mathfrak{D}$ and

$$\lim_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x}) = u, \qquad \lim_{\mathbf{x}\to\mathbf{x}_0} g(\mathbf{x}) = v,$$

then

$$\lim_{\mathbf{x}\to\mathbf{x}_0} (fg)(\mathbf{x}) = uv.$$

If $g(\mathbf{x}) \neq 0$ for all $\mathbf{x} \in \mathfrak{D}$, and $v \neq 0$, then

$$\lim_{\mathbf{x}\to\mathbf{x}_0} \left(\frac{f}{g}\right)(\mathbf{x}) = \frac{u}{v}.$$

Example 2.12

If $\mathbf{k} = (k_1, \ldots, k_n)$ is a k-tuple of nonnegative integers, the monomial $p_{\mathbf{k}} : \mathbb{R}^n \to \mathbb{R}$,

$$p_{\mathbf{k}}(\mathbf{x}) = x_1^{k_1} \cdots x_n^{k_n}$$

can be written as a product of the projection functions $\pi_i : \mathbb{R}^n \to \mathbb{R}$, $\pi_i(\mathbf{x}) = x_i$, $1 \leq i \leq n$. By Proposition 2.12,

$$\lim_{\mathbf{x}\to\mathbf{x}_0} p_{\mathbf{k}}(\mathbf{x}) = p_{\mathbf{k}}(\mathbf{x}_0)$$

for any $\mathbf{x}_0$ in $\mathbb{R}^n$. If $p : \mathbb{R}^n \to \mathbb{R}$ is a polynomial, it is a finite linear combination of monomials. Proposition 2.11 then implies that for any $\mathbf{x}_0$ in $\mathbb{R}^n$,

$$\lim_{\mathbf{x}\to\mathbf{x}_0} p(\mathbf{x}) = p(\mathbf{x}_0).$$

If $f : \mathfrak{D} \to \mathbb{R}$, $f(\mathbf{x}) = p(\mathbf{x})/q(\mathbf{x})$ is a rational function which is equal to the quotient of the polynomial $p(\mathbf{x})$ by the polynomial $q(\mathbf{x})$, then Proposition 2.12 implies that

$$\lim_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x}) = f(\mathbf{x}_0)$$

for any $\mathbf{x}_0 \in \mathfrak{D} = \{\mathbf{x} \in \mathbb{R}^n \,|\, q(\mathbf{x}) \neq 0\}$.

Example 2.13

Find $\lim\limits_{(x,y)\to(1,-1)} \dfrac{x^2+3xy+2y^2}{x^2+y^2}$.

Solution

Since

$$\lim_{(x,y)\to(1,-1)} (x^2+3xy+2y^2) = 1-3+2 = 0,$$
$$\lim_{(x,y)\to(1,-1)} (x^2+y^2) = 1+1 = 2,$$

we find that

$$\lim_{(x,y)\to(1,-1)} \frac{x^2+3xy+2y^2}{x^2+y^2} = \frac{0}{2} = 0.$$

It is easy to deduce the limit law for composite functions.

Proposition 2.13

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$, and let $\mathcal{U}$ be a subset of $\mathbb{R}^k$. Given the two functions $\mathbf{F}:\mathfrak{D}\to\mathbb{R}^k$ and $\mathbf{G}:\mathcal{U}\to\mathbb{R}^m$, if $\mathbf{F}(\mathfrak{D})\subset\mathcal{U}$, we can define the composite function $\mathbf{H}=\mathbf{G}\circ\mathbf{F}:\mathfrak{D}\to\mathbb{R}^m$ by $\mathbf{H}(\mathbf{x})=\mathbf{G}(\mathbf{F}(\mathbf{x}))$. If $\mathbf{x}_0$ is a limit point of $\mathfrak{D}$, $\mathbf{y}_0$ is a limit point of $\mathcal{U}$, $\mathbf{F}(\mathfrak{D}\setminus\{\mathbf{x}_0\})\subset\mathcal{U}\setminus\{\mathbf{y}_0\}$,

$$\lim_{\mathbf{x}\to\mathbf{x}_0}\mathbf{F}(\mathbf{x})=\mathbf{y}_0, \qquad \lim_{\mathbf{y}\to\mathbf{y}_0}\mathbf{G}(\mathbf{y})=\mathbf{v},$$

then

$$\lim_{\mathbf{x}\to\mathbf{x}_0}\mathbf{H}(\mathbf{x}) = \lim_{\mathbf{x}\to\mathbf{x}_0}(\mathbf{G}\circ\mathbf{F})(\mathbf{x}) = \mathbf{v}.$$

The proof repeats verbatim the proof of the corresponding theorem for single variable functions.

Example 2.14

Find the limit $\lim\limits_{(x,y)\to(0,0)} \dfrac{\sin(2x^2+3y^2)}{2x^2+3y^2}$.

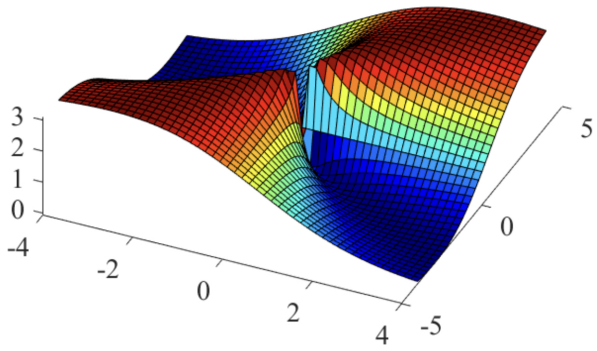


Figure 2.1: The function $f(x,y) = \dfrac{x^2+3xy+2y^2}{x^2+y^2}$ in Example 2.13.

Solution

Since

$$\lim_{(x,y)\to(0,0)} (2x^2+3y^2) = 2\times 0 + 3\times 0 = 0, \qquad \lim_{u\to 0} \frac{\sin u}{u} = 1,$$

the limit law for composite functions implies that

$$\lim_{(x,y)\to(0,0)} \frac{\sin(2x^2+3y^2)}{2x^2+3y^2} = 1.$$

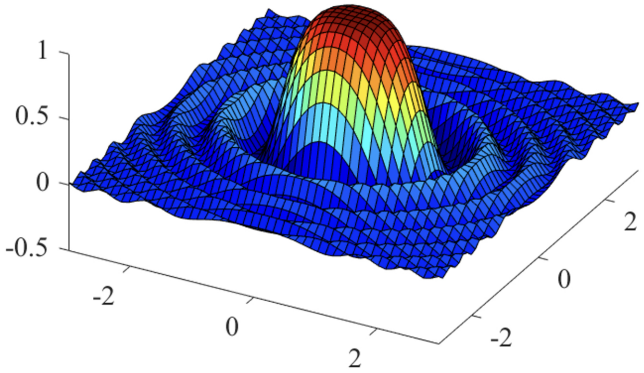


Figure 2.2: The function $f(x,y) = \dfrac{\sin(2x^2+3y^2)}{2x^2+3y^2}$ in Example 2.14.

Let us look at some examples where the rules we have studied cannot be applied.

Example 2.15

Determine whether the limit $\lim\limits_{(x,y)\to(0,0)} \dfrac{x^2-2y^2}{x^2+y^2}$ exists.

Solution

Let

$$f(x,y)=\frac{x^2-2y^2}{x^2+y^2}=\frac{p(x,y)}{q(x,y)}.$$

When $(x,y)\to(0,0)$, $q(x,y)=x^2+y^2\to 0$. Hence, we cannot apply limit law for quotients of functions.

Consider the sequences of points $\{\mathbf{u}_k\}$ and $\{\mathbf{v}_k\}$ in $\mathbb{R}^2\setminus\{0,0\}$ given by

$$\mathbf{u}_k=\left(\frac{1}{k},0\right),\qquad \mathbf{v}_k=\left(0,\frac{1}{k}\right).$$

Notice that both the sequences $\{\mathbf{u}_k\}$ and $\{\mathbf{v}_k\}$ converge to $(0,0)$. If $\lim\limits_{(x,y)\to(0,0)} f(x,y)=a$, then both the sequences $\{f(\mathbf{u}_k)\}$ and $\{f(\mathbf{v}_k)\}$ should converge to a. Since

$$f(\mathbf{u}_k)=1,\qquad f(\mathbf{v}_k)=-2\qquad \text{for all } k\in\mathbb{Z}^+,$$

the sequence $\{f(\mathbf{u}_k)\}$ converges to 1, while the sequence $\{f(\mathbf{v}_k)\}$ converges to -2. These imply that $a=1$ and $a=-2$, which is a contradiction. Hence, the limit $\lim\limits_{(x,y)\to(0,0)} \dfrac{x^2-2y^2}{x^2+y^2}$ does not exist.

Example 2.16

Determine whether the limit $\lim\limits_{(x,y)\to(0,0)} \dfrac{xy}{x^2+2y^2}$ exists.

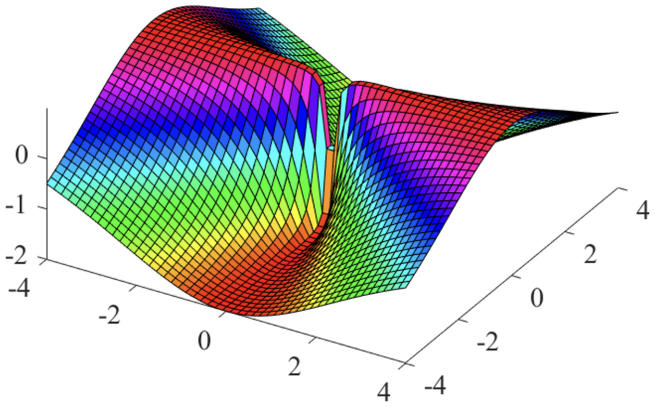


Figure 2.3: The function $f(x,y) = \dfrac{x^2 - 2y^2}{x^2 + y^2}$ in Example 2.15.

Solution

Let

$$f(x,y) = \frac{xy}{x^2 + 2y^2}.$$

Consider the sequences of points $\{\mathbf{u}_k\}$ and $\{\mathbf{v}_k\}$ in $\mathbb{R}^2 \setminus \{0,0\}$ given by

$$\mathbf{u}_k = \left(\frac{1}{k}, 0\right), \qquad \mathbf{v}_k = \left(\frac{1}{k}, \frac{1}{k}\right),$$

Notice that both the sequences $\{\mathbf{u}_k\}$ and $\{\mathbf{v}_k\}$ converge to $(0,0)$. If $\lim\limits_{(x,y)\to(0,0)} f(x,y) = a$, then both the sequences $\{f(\mathbf{u}_k)\}$ and $\{f(\mathbf{v}_k)\}$ should converge to a. Since

$$f(\mathbf{u}_k) = 0, \qquad f(\mathbf{v}_k) = \frac{1}{3} \qquad \text{for all } k \in \mathbb{Z}^+,$$

the sequence $\{f(\mathbf{u}_k)\}$ converges to 0, while the sequence $\{f(\mathbf{v}_k)\}$ converges to $1/3$. These imply that $a = 0$ and $a = 1/3$, which is a contradiction. Hence, the limit $\lim\limits_{(x,y)\to(0,0)} \dfrac{xy}{x^2 + 2y^2}$ does not exist.

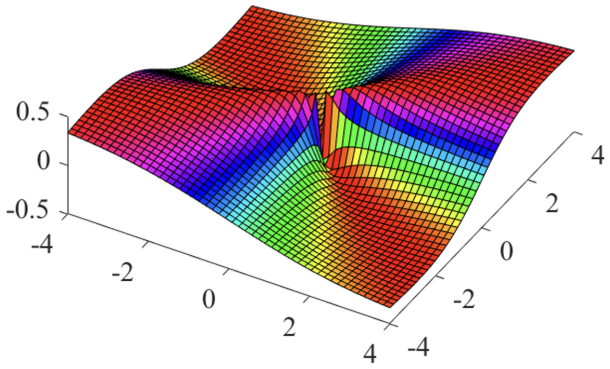


Figure 2.4: The function $f(x,y) = \dfrac{xy}{x^2+2y^2}$ in Example 2.16.

Example 2.17

Determine whether the limit $\lim\limits_{(x,y)\to(0,0)} \dfrac{xy^2}{x^2+2y^4}$ exists.

Solution

Let

$$f(x,y) = \frac{xy^2}{x^2+2y^4}.$$

Consider the sequences of points $\{\mathbf{u}_k\}$ and $\{\mathbf{v}_k\}$ in $\mathbb{R}^2 \setminus \{0,0\}$ given by

$$\mathbf{u}_k = \left(\frac{1}{k}, 0\right), \qquad \mathbf{v}_k = \left(\frac{1}{k^2}, \frac{1}{k}\right),$$

Notice that both the sequences $\{\mathbf{u}_k\}$ and $\{\mathbf{v}_k\}$ converge to $(0,0)$. If $\lim\limits_{(x,y)\to(0,0)} f(x,y) = a$, then both the sequences $\{f(\mathbf{u}_k)\}$ and $\{f(\mathbf{v}_k)\}$ should converge to a. Since

$$f(\mathbf{u}_k) = 0, \qquad f(\mathbf{v}_k) = \frac{1}{3} \qquad \text{for all } k \in \mathbb{Z}^+,$$

the sequence $\{f(\mathbf{u}_k)\}$ converges to 0, while the sequence $\{f(\mathbf{v}_k)\}$ converges to $1/3$. These imply that $a = 0$ and $a = 1/3$, which is a contradiction. Hence, the limit $\lim\limits_{(x,y)\to(0,0)} \dfrac{xy^2}{x^2+2y^4}$ does not exist.

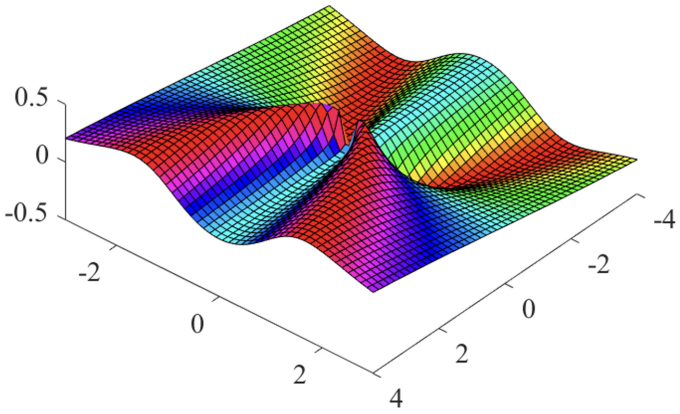


Figure 2.5: The function $f(x,y) = \dfrac{xy^2}{x^2+2y^4}$ in Example 2.17.

Example 2.18

Determine whether the limit $\lim\limits_{(x,y)\to(0,0)} \dfrac{xy^2}{x^2+2y^2}$ exists.

Solution

Let

$$f(x,y) = \frac{xy^2}{x^2+2y^2}.$$

If $\{(x_k,y_k)\}$ is a sequence of points in $\mathbb{R}^2 \setminus \{0,0\}$ that converges to $(0,0)$, then

$$|f(x_k,y_k)| = |x_k|\frac{y_k^2}{x_k^2+2y_k^2} \le |x_k|.$$

The sequence $\{x_k\}$ converges to 0. By squeeze theorem, the sequence $\{f(x_k,y_k)\}$ also converges to 0. This proves that

$$\lim_{(x,y)\to(0,0)} \frac{xy^2}{x^2+2y^2} = 0.$$

Similar to the single variable case, there is an equivalent definition of limits in terms of ε and δ.

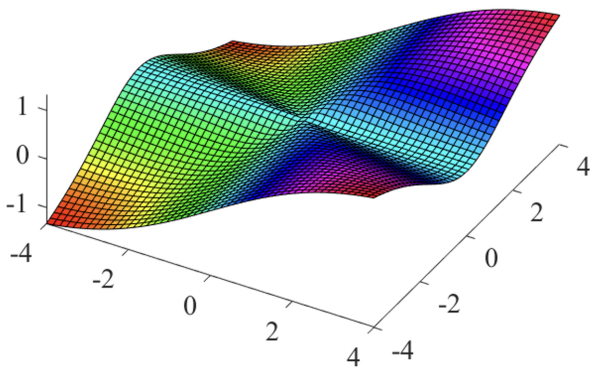


Figure 2.6: The function $f(x, y) = \dfrac{xy^2}{x^2 + 2y^2}$ in Example 2.18.

Theorem 2.14 Equivalent Definitions for Limits

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$, and let $\mathbf{x}_0$ be a limit point of $\mathfrak{D}$. Given a function $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$, the following two definitions for

$$\lim_{\mathbf{x} \to \mathbf{x}_0} \mathbf{F}(\mathbf{x}) = \mathbf{v}$$

are equivalent.

(i) Whenever $\{\mathbf{x}_k\}$ is a sequence of points in $\mathfrak{D} \setminus \{\mathbf{x}_0\}$ that converges to $\mathbf{x}_0$, the sequence $\{\mathbf{F}(\mathbf{x}_k)\}$ converges to $\mathbf{v}$.

(ii) For any $\varepsilon > 0$, there is a $\delta > 0$ such that if the point $\mathbf{x}$ is in $\mathfrak{D}$ and $0 < \|\mathbf{x} - \mathbf{x}_0\| < \delta$, then $\|\mathbf{F}(\mathbf{x}) - \mathbf{v}\| < \varepsilon$.

Proof

We will prove that if (ii) holds, then (i) holds; and if (ii) does not hold, then (i) also does not hold.

First assume that (ii) holds. If $\{\mathbf{x}_k\}$ is a sequence in $\mathfrak{D} \setminus \{\mathbf{x}_0\}$ that converges to the point $\mathbf{x}_0$, we need to show that the sequence $\{\mathbf{F}(\mathbf{x}_k)\}$ converges to $\mathbf{v}$. Given $\varepsilon > 0$, (ii) implies that there is a $\delta > 0$ such that for all $\mathbf{x}$ that is in $\mathfrak{D} \setminus \{\mathbf{x}_0\}$ with $\|\mathbf{x} - \mathbf{x}_0\| < \delta$, we have $\|\mathbf{F}(\mathbf{x}) - \mathbf{v}\| < \varepsilon$.

Since $\{\mathbf{x}_k\}$ converges to $\mathbf{x}_0$, there is a positive integer K such that for all $k \geq K$, $\|\mathbf{x}_k - \mathbf{x}_0\| < \delta$. Therefore, for all $k \geq K$, $\|\mathbf{F}(\mathbf{x}_k) - \mathbf{v}\| < \varepsilon$. This shows that the sequence $\{\mathbf{F}(\mathbf{x}_k)\}$ indeed converges to $\mathbf{v}$.
Now assume that (ii) does not hold. Then there is an $\varepsilon > 0$ such that for any $\delta > 0$, there is a point $\mathbf{x}$ in $\mathfrak{D}\setminus\{\mathbf{x}_0\}$ with $\|\mathbf{x}-\mathbf{x}_0\| < \delta$ but $\|\mathbf{F}(\mathbf{x}_k)-\mathbf{v}\| \geq \varepsilon$. For this $\varepsilon > 0$, we construct a sequence $\{\mathbf{x}_k\}$ in $\mathfrak{D} \setminus \{\mathbf{x}_0\}$ in the following way. For each positive integer k, there is a point $\mathbf{x}_k$ in $\mathfrak{D} \setminus \{\mathbf{x}_0\}$ such that $\|\mathbf{x}-\mathbf{x}_0\| < 1/k$ but $\|\mathbf{F}(\mathbf{x}_k)-\mathbf{v}\| \geq \varepsilon$. Then $\{\mathbf{x}_k\}$ is a sequence in $\mathfrak{D}\setminus\{\mathbf{x}_0\}$ that satisfies

$$\|\mathbf{x} - \mathbf{x}_0\| < 1/k \qquad \text{for all } k \in \mathbb{Z}^+.$$

Hence, it converges to $\mathbf{x}_0$. Since $\|\mathbf{F}(\mathbf{x}_k) - \mathbf{v}\| \geq \varepsilon$ for all $k \in \mathbb{Z}^+$, the sequence $\{\mathbf{F}(\mathbf{x}_k)\}$ cannot converge to $\mathbf{v}$. This proves that (i) does not hold.

We can give an alternative solution to Example 2.18 as follows.

Alternative Solution to Example 2.18

Let

$$f(x, y) = \frac{xy^2}{x^2 + 2y^2}.$$

Given $\varepsilon > 0$, let $\delta = \varepsilon$. If (x, y) is a point in $\mathbb{R}^2 \setminus \{(0, 0)\}$ such that

$$\sqrt{x^2 + y^2} = \|(x, y) - (0, 0)\| < \delta = \varepsilon,$$

then $|x| < \varepsilon$. This implies that

$$|f(x, y) - 0| = |x|\frac{y^2}{x^2 + 2y^2} \leq |x| < \varepsilon.$$

Hence,

$$\lim_{(x,y)\to(0,0)} \frac{xy^2}{x^2 + 2y^2} = 0.$$

Exercises 2.2

Question 1

Determine whether the limit exists. If it exists, find the limit.

(a) $\lim\limits_{(x,y)\to(1,2)} \dfrac{4x^2 - y^2}{x^2 + y^2}$

(b) $\lim\limits_{(x,y)\to(1,2)} \sqrt{\dfrac{4x^2 - y^2}{x^2 + y^2}}$

(c) $\lim\limits_{(x,y)\to(1,2)} \sqrt{\dfrac{4x^2 + y^2}{x^2 + y^2}}$

Question 2

Determine whether the limit exists. If it exists, find the limit.

(a) $\lim\limits_{(x,y)\to(0,0)} \dfrac{x^3 + y^3}{x^2 + y^2}$

(b) $\lim\limits_{(x,y)\to(0,0)} \dfrac{x^2 + y^3}{x^2 + y^2}$

(c) $\lim\limits_{(x,y)\to(0,0)} \dfrac{e^{4x^2+y^2} - 1}{4x^2 + y^2}$

(d) $\lim\limits_{(x,y)\to(0,0)} \dfrac{e^{x^2+y^2} - 1}{4x^2 + y^2}$

Question 3

Determine whether the limit

$$\lim_{(x,y)\to(0,0)} \frac{x^2 + 4y^4}{4x^2 + y^4}$$

exists. If it exists, find the limit.

Question 4

Determine whether the limit

$$\lim_{(x,y)\to(1,1)} \frac{\cos(x^2+y^2-2)-1}{(x^2+y^2-2)^2}$$

exists. If it exists, find the limit.

Question 5

Let $\mathbf{x}_0$ be a point in $\mathbb{R}^n$. Find the limit $\lim_{\mathbf{x}\to\mathbf{x}_0} \dfrac{\mathbf{x}}{\|\mathbf{x}\|}$.

Question 6

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$, and let $f : \mathfrak{D} \to \mathbb{R}$ and $\mathbf{G} : \mathfrak{D} \to \mathbb{R}^m$ be functions defined on $\mathfrak{D}$. We can define the function $\mathbf{H} : \mathfrak{D} \to \mathbb{R}^m$ by

$$\mathbf{H}(\mathbf{x}) = f(\mathbf{x})\mathbf{G}(\mathbf{x}) \qquad \text{for all } \mathbf{x} \in \mathfrak{D}.$$

If $\mathbf{x}_0$ is a point in $\mathfrak{D}$ and

$$\lim_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x}) = a, \qquad \lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{G}(\mathbf{x}) = \mathbf{v},$$

show that

$$\lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{H}(\mathbf{x}) = a\mathbf{v}.$$

2.3 Continuity

The definition of continuity is a direct generalization of the single variable case.

Definition 2.7 Continuity

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ be a function defined on $\mathfrak{D}$. We say that the function $\mathbf{F}$ is **continuous at** $\mathbf{x}_0$ provided that whenever $\{\mathbf{x}_k\}$ is a sequence of points in $\mathfrak{D}$ that converges to $\mathbf{x}_0$, the sequence $\{\mathbf{F}(\mathbf{x}_k)\}$ converges to $\mathbf{F}(\mathbf{x}_0)$.
We say that $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is a **continuous function** if it is continuous at every point of its domain $\mathfrak{D}$.

From the definition, we obtain the following immediately.

Proposition 2.15 Limits and Continuity

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ be a function defined on $\mathfrak{D}$.

1. If $\mathbf{x}_0$ is an isolated point of $\mathfrak{D}$, then $\mathbf{F}$ is continuous at $\mathbf{x}_0$.

2. If $\mathbf{x}_0$ is a limit point of $\mathfrak{D}$, then $\mathbf{F}$ is continuous at $\mathbf{x}_0$ if and only if

$$\lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{F}(\mathbf{x}) = \mathbf{F}(\mathbf{x}_0).$$

Example 2.19

Example 2.10 says that for each $1 \leq i \leq n$, the projection function $\pi_i : \mathbb{R}^n \to \mathbb{R}$, $\pi_i(\mathbf{x}) = x_i$, is a continuous function.

Example 2.20

Example 2.11 says that the norm function $f : \mathbb{R}^n \to \mathbb{R}$, $f(\mathbf{x}) = \|\mathbf{x}\|$, is a continuous function.

From Proposition 2.10, we have the following.

Proposition 2.16

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ be a function defined on $\mathfrak{D}$. The function $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is continuous at $\mathbf{x}_0$ if and only if each of the component functions $F_j = (\pi_j \circ \mathbf{F}) : \mathfrak{D} \to \mathbb{R}$, $1 \leq j \leq m$, is continuous at $\mathbf{x}_0$.

Example 2.21

The function $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^2$,

$$\mathbf{F}(x, y, z) = (x, z),$$

is a continuous function since each component function is continuous.

Proposition 2.11 gives the following.

Proposition 2.17

Let $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ and $\mathbf{G} : \mathfrak{D} \to \mathbb{R}^m$ be functions defined on $\mathfrak{D} \subset \mathbb{R}^n$, and let $\mathbf{x}_0$ be a point in $\mathfrak{D}$. If $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ and $\mathbf{G} : \mathfrak{D} \to \mathbb{R}^m$ are continuous at $\mathbf{x}_0$, then for any real numbers α and β, the function $(\alpha\mathbf{F} + \beta\mathbf{G}) : \mathfrak{D} \to \mathbb{R}^m$ is continuous at $\mathbf{x}_0$.

Proposition 2.12 gives the following.

Proposition 2.18

Let $f : \mathfrak{D} \to \mathbb{R}$ and $g : \mathfrak{D} \to \mathbb{R}$ be functions defined on $\mathfrak{D} \subset \mathbb{R}^n$, and let $\mathbf{x}_0$ be a point in $\mathfrak{D}$. Assume that the functions $f : \mathfrak{D} \to \mathbb{R}$ and $g : \mathfrak{D} \to \mathbb{R}$ are continuous at $\mathbf{x}_0$.

1. The function $(fg) : \mathfrak{D} \to \mathbb{R}$ is continuous at $\mathbf{x}_0$.
2. If $g(\mathbf{x}) \neq 0$ for all $\mathbf{x} \in \mathfrak{D}$, then the function $(f/g) : \mathfrak{D} \to \mathbb{R}$ is continuous at $\mathbf{x}_0$.

Example 2.12 gives the following.

Proposition 2.19

Polynomials and rational functions are continuous functions.

Since each component of a linear transformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^m$ is a polynomial, we have the following.

Proposition 2.20

A linear transformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^m$ is a continuous function.

Since a quadratic form $Q : \mathbb{R}^n \to \mathbb{R}$ is a polynomial, we have the following.

Proposition 2.21

A quadratic form $Q : \mathbb{R}^n \to \mathbb{R}$ given by

$$Q(\mathbf{x}) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$$

is a continuous function.

The following is obvious from the definition of continuity.

Proposition 2.22

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$, and let $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ be a function that is continuous at the point $\mathbf{x}_0 \in \mathfrak{D}$. If $\mathfrak{D}_1$ is a subset of $\mathfrak{D}$ that contains $\mathbf{x}_0$, then the function $\mathbf{F} : \mathfrak{D}_1 \to \mathbb{R}^m$ is also continuous at $\mathbf{x}_0$.

Example 2.22

Let $\mathfrak{D}$ be the set

$$\mathfrak{D} = \left\{ (x, y) \,|\, x^2 + y^2 < 1 \right\},$$

and let $f : \mathfrak{D} \to \mathbb{R}$ be the function defined as

$$f(x, y) = \frac{xy}{1 - x^2 - y^2}.$$

Since $f_1(x, y) = xy$ and $f_2(x, y) = 1 - x^2 - y^2$ are polynomials, they are continuous. Since $f_2(x, y) \neq 0$ for all $(x, y) \in \mathfrak{D}$, $f : \mathfrak{D} \to \mathbb{R}$ is a continuous function.

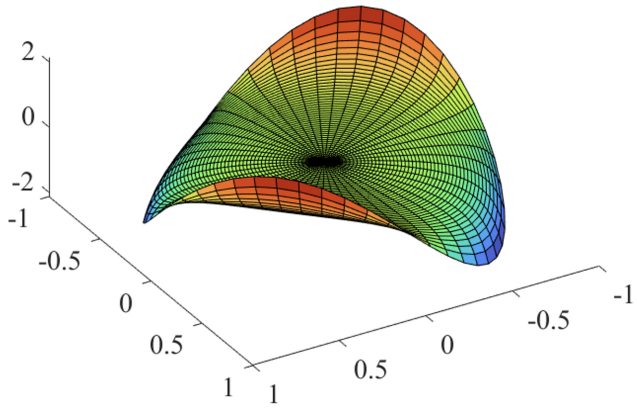


Figure 2.7: The function $f(x, y) = \dfrac{xy}{1 - x^2 - y^2}$ in Example 2.22.

Proposition 2.13 implies the following.

Proposition 2.23

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$, and let $\mathcal{U}$ be a subset of $\mathbb{R}^k$. If $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^k$ and $\mathbf{G} : \mathcal{U} \to \mathbb{R}^m$ are functions such that $\mathbf{F}(\mathfrak{D}) \subset \mathcal{U}$, $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^k$ is continuous at $\mathbf{x}_0$, $\mathbf{G} : \mathcal{U} \to \mathbb{R}^m$ is continuous at $\mathbf{y}_0$, then the composite function $\mathbf{H} = (\mathbf{G} \circ \mathbf{F}) : \mathfrak{D} \to \mathbb{R}^m$ is continuous at $\mathbf{x}_0$.

A direct proof of this theorem using the definition of continuity is actually much simpler.

Proof

If $\{\mathbf{x}_k\}$ is a sequence of points in $\mathfrak{D}$ that converges to $\mathbf{x}_0$, then since $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^k$ is continuous at $\mathbf{x}_0$, $\{\mathbf{F}(\mathbf{x}_k)\}$ is a sequence of points in $\mathcal{U}$ that converges to $\mathbf{y}_0$. Since $\mathbf{G} : \mathcal{U} \to \mathbb{R}^m$ is continuous at $\mathbf{y}_0$, $\{\mathbf{G}(\mathbf{F}(\mathbf{x}_k))\}$ is a sequence of points in $\mathbb{R}^m$ that converges to $\mathbf{G}(\mathbf{y}_0) = \mathbf{G}(\mathbf{F}(\mathbf{x}_0))$.

In other words, the sequence $\{\mathbf{H}(\mathbf{x}_k)\}$ converges to $\mathbf{H}(\mathbf{x}_0)$. This shows that the function $\mathbf{H} = (\mathbf{G} \circ \mathbf{F}) : \mathfrak{D} \to \mathbb{R}^m$ is continuous at $\mathbf{x}_0$.

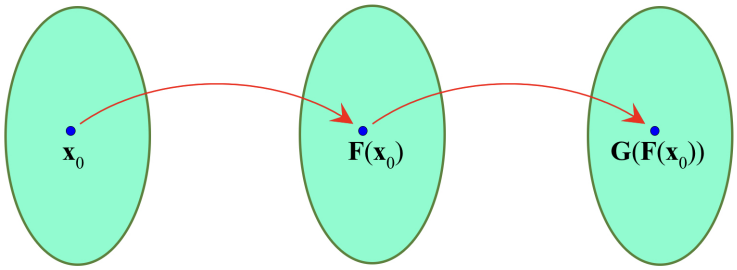


Figure 2.8: Composition of functions.

Corollary 2.24

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$, and let $\mathbf{x}_0$ be a point in $\mathfrak{D}$. If the function $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is continuous at $\mathbf{x}_0 \in \mathfrak{D}$, then the function $\|\mathbf{F}\| : \mathfrak{D} \to \mathbb{R}$ is also continuous at $\mathbf{x}_0$.

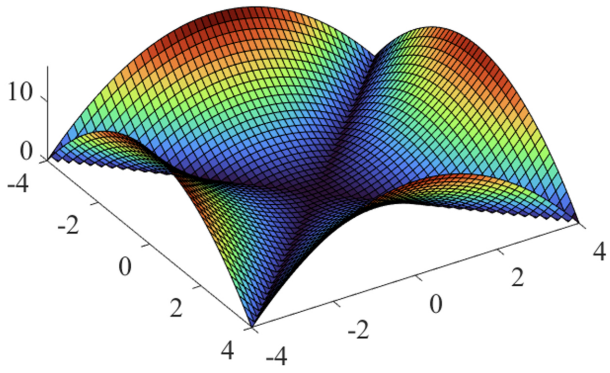


Figure 2.9: The function $f(x, y) = |x^2 - y^2|$.

Example 2.23

The function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = |x^2 - y^2|$ is a continuous function since $f(x, y) = |p(x, y)|$, where $p(x, y) = x^2 - y^2$ is a polynomial function, which is continuous.

Example 2.24

Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = \sqrt{e^{2xy} + x^2 + y^2}$. Notice that $f(x, y) = \|\mathbf{F}(x, y)\|$, where $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^3$ is the function given by

$$\mathbf{F}(x, y) = (e^{xy}, x, y)\,.$$

Since $g(x, y) = xy$ is a polynomial function, it is continuous. Being a composition of the continuous function $h(x) = e^x$ with the continuous function $g(x, y) = xy$, $F_1(x, y) = (h \circ g)(x, y) = e^{xy}$ is a continuous function. The functions $F_2(x, y) = x$ and $F_3(x, y) = y$ are continuous functions. Hence, $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^3$ is a continuous function. This implies that $f : \mathbb{R}^2 \to \mathbb{R}$ is also a continuous function.

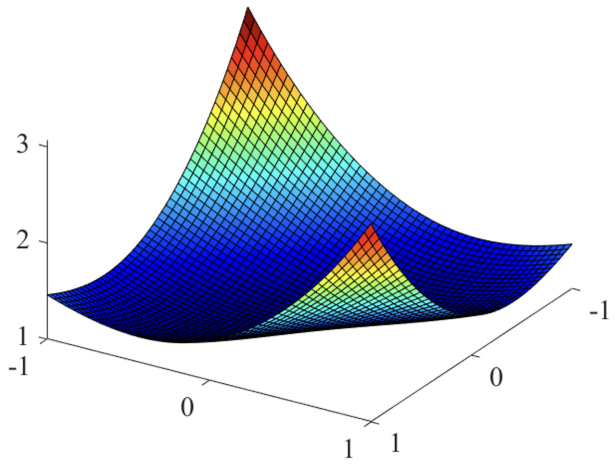


Figure 2.10: The function $f(x, y) = \sqrt{e^{2xy} + x^2 + y^2}$.

Example 2.25

We have shown in volume I that the function $f : \mathbb{R} \to \mathbb{R}$,

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

is a continuous function. Define the function $h : \mathbb{R}^3 \to \mathbb{R}$ by

$$h(x, y, z) = \begin{cases} \dfrac{\sin(x^2 + y^2 + z^2)}{x^2 + y^2 + z^2}, & \text{if } (x, y, z) \neq (0, 0, 0), \\ 1, & \text{if } (x, y, z) = (0, 0, 0). \end{cases}$$

Since $h = f \circ g$, where $g : \mathbb{R}^3 \to \mathbb{R}$ is the polynomial function $g(x, y, z) = x^2 + y^2 + z^2$, which is continuous, the function $h : \mathbb{R}^3 \to \mathbb{R}$ is continuous.

The following gives an equivalent definition of continuity in terms of ε and δ.

Theorem 2.25 Equivalent Definitions of Continuity

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$, and let $\mathbf{x}_0$ be a limit point of $\mathfrak{D}$. Given a function $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$, the following two definitions for the continuity of $\mathbf{F}$ at $\mathbf{x}_0$ are equivalent.

(i) Whenever $\{\mathbf{x}_k\}$ is a sequence of points in $\mathfrak{D}$ that converges to $\mathbf{x}_0$, the sequence $\{\mathbf{F}(\mathbf{x}_k)\}$ converges to $\mathbf{F}(\mathbf{x}_0)$.

(ii) For any $\varepsilon > 0$, there is a $\delta > 0$ such that if the point $\mathbf{x}$ is in $\mathfrak{D}$ and $\|\mathbf{x} - \mathbf{x}_0\| < \delta$, then $\|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}_0)\| < \varepsilon$.

The proof is left as an exercise. Notice that statement (ii) can be reformulated as follows. For any $\varepsilon > 0$, there is a $\delta > 0$ such that if the point $\mathbf{x}$ is in $\mathfrak{D}$ and $\mathbf{x} \in B(\mathbf{x}_0, \delta)$, then $\mathbf{F}(\mathbf{x}) \in B(\mathbf{F}(\mathbf{x}_0), \varepsilon)$.

Now we want to explore another important property of continuity.

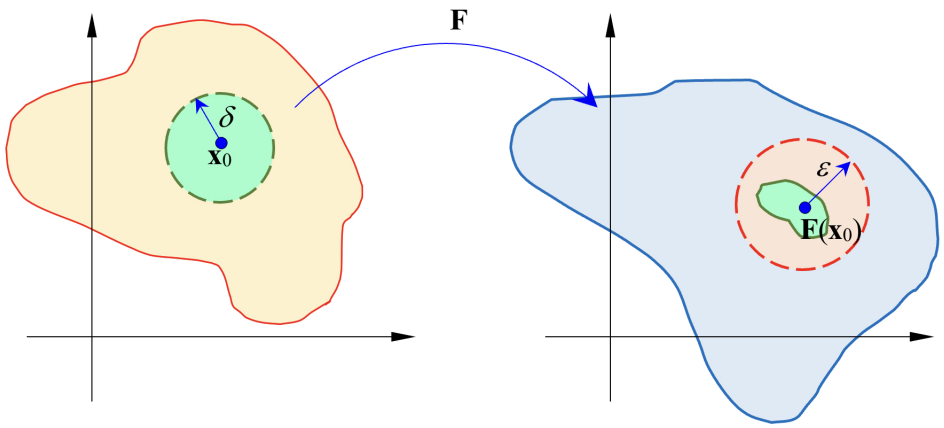


Figure 2.11: The definition of continuity in terms of ε and δ.

Theorem 2.26

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ be a function defined on $\mathcal{O}$. The following are equivalent.

(a) $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is continuous.

(b) For every open subset V of $\mathbb{R}^m$, $\mathbf{F}^{-1}(V)$ is an open subset of $\mathbb{R}^n$.

Note that for this theorem to hold, it is important that the domain of the function $\mathbf{F}$ is an open set.

Proof

Assume that (a) holds. Let V be an open subset of $\mathbb{R}^m$, and let

$$U = \mathbf{F}^{-1}(V) = \{\mathbf{x} \in \mathcal{O} \,|\, \mathbf{F}(\mathbf{x}) \in V\}\,.$$

We need to show that U is an open subset of $\mathbb{R}^n$. If $\mathbf{x}_0$ is in U, then it is in $\mathcal{O}$. Since $\mathcal{O}$ is open, there exists $r_0 > 0$ such that $B(\mathbf{x}_0, r_0) \subset \mathcal{O}$. Since $\mathbf{y}_0 = \mathbf{F}(\mathbf{x}_0)$ is in V and V is open, there exists $\varepsilon > 0$ such that $B(\mathbf{y}_0, \varepsilon) \subset V$. By (a), there exists $\delta > 0$ such that for any $\mathbf{x} \in \mathcal{O}$, if $\|\mathbf{x} - \mathbf{x}_0\| < \delta$, then $\|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}_0)\| < \varepsilon$.

Take $r = \min\{\delta, r_0\}$. Then $r > 0$, $r \leq r_0$ and $r \leq \delta$. If $\mathbf{x}$ is in $B(\mathbf{x}_0, r)$, then $\mathbf{x} \in \mathcal{O}$ and $\|\mathbf{x} - \mathbf{x}_0\| < r \leq \delta$. It follows that $\|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}_0)\| < \varepsilon$. This implies that $\mathbf{F}(\mathbf{x}) \in B(\mathbf{y}_0, \varepsilon) \subset V$. Thus, $\mathbf{x} \in U$. In other words, we have shown that $B(\mathbf{x}_0, r)$ is contained in U. This proves that U is open, which is the assertion of (b).

Conversely, assume that (b) holds. Let $\mathbf{x}_0$ be a point in $\mathcal{O}$, and let $\mathbf{y}_0 = \mathbf{F}(\mathbf{x}_0)$. Given $\varepsilon > 0$, the ball $V = B(\mathbf{y}_0, \varepsilon)$ is an open subset of $\mathbb{R}^m$. By (b), $U = \mathbf{F}^{-1}(V)$ is open in $\mathbb{R}^n$. By definition, U is a subset of $\mathcal{O}$. Since $\mathbf{F}(\mathbf{x}_0)$ is in V, $\mathbf{x}_0$ is in U. Since U is open and it contains $\mathbf{x}_0$, there is an $r > 0$ such that $B(\mathbf{x}_0, r) \subset U$. Take $\delta = r$. Then if $\mathbf{x}$ is a point in $\mathcal{O}$ and $\|\mathbf{x} - \mathbf{x}_0\| < r$, $\mathbf{x} \in B(\mathbf{x}_0, r) \subset U$. This implies that $\mathbf{F}(\mathbf{x}) \in V = B(\mathbf{y}_0, \varepsilon)$. Namely, $\|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}_0)\| < \varepsilon$. This proves that $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is continuous at $\mathbf{x}_0$. Since $\mathbf{x}_0$ is an arbitrary point in $\mathcal{O}$, $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is continuous.

Using the fact that a set is open if and only if its complement is closed, it is natural to expect the following.

Theorem 2.27

Let $\mathcal{A}$ be a closed subset of $\mathbb{R}^n$, and let $\mathbf{F} : \mathcal{A} \to \mathbb{R}^m$ be a function defined on $\mathcal{A}$. The following are equivalent.

(a) $\mathbf{F} : \mathcal{A} \to \mathbb{R}^m$ is continuous.

(b) For every closed subset C of $\mathbb{R}^m$, $\mathbf{F}^{-1}(C)$ is a closed subset of $\mathbb{R}^n$.

Proof

Assume that (a) holds. Let C be a closed subset of $\mathbb{R}^m$, and let

$$D = \mathbf{F}^{-1}(C) = \{\mathbf{x} \in \mathcal{A} \,|\, \mathbf{F}(\mathbf{x}) \in D\}\,.$$

We need to show that D is a closed subset of $\mathbb{R}^n$. If $\{\mathbf{x}_k\}$ is a sequence in D that converges to the point $\mathbf{x}_0$ in $\mathbb{R}^n$, since $D \subset \mathcal{A}$ and $\mathcal{A}$ is closed, $\mathbf{x}_0$ is in $\mathcal{A}$. Since $\mathbf{F}$ is continuous at $\mathbf{x}_0$, the sequence $\{\mathbf{F}(\mathbf{x}_k)\}$ is a sequence in C that converges to the point $\mathbf{F}(\mathbf{x}_0)$ in $\mathbb{R}^m$. Since C is closed, $\mathbf{F}(\mathbf{x}_0)$ is in C. Therefore, $\mathbf{x}_0$ is in D. This proves that D is closed.

Conversely, assume that (a) does not hold. Then $\mathbf{F} : \mathcal{A} \to \mathbb{R}^m$ is not continuous at some $\mathbf{x}_0 \in \mathcal{A}$. Thus, there exists $\varepsilon > 0$ such that for any $\delta > 0$, there exists a point $\mathbf{x}$ in $\mathcal{A} \cap B(\mathbf{x}_0, \delta)$ such that $\|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}_0)\| \geq \varepsilon$. For $k \in \mathbb{Z}^+$, let $\mathbf{x}_k$ be a point in $\mathcal{A} \cap B(\mathbf{x}_0, 1/k)$ such that $\|\mathbf{F}(\mathbf{x}_k) - \mathbf{F}(\mathbf{x}_0)\| \geq \varepsilon$. Since

$$\|\mathbf{x}_k - \mathbf{x}_0\| < \frac{1}{k} \qquad \text{for all } k \in \mathbb{Z}^+,$$

the sequence $\{\mathbf{x}_k\}$ is a sequence in $\mathcal{A}$ that converges to $\mathbf{x}_0$. Let

$$C = \{\mathbf{y} \in \mathbb{R}^m \,|\, \|\mathbf{y} - \mathbf{F}(\mathbf{x}_0)\| \geq \varepsilon\}.$$

Then C is the complement of the open set $B(\mathbf{F}(\mathbf{x}_0), \varepsilon)$. Hence, C is closed. It contains $\mathbf{F}(\mathbf{x}_k)$ for all $k \in \mathbb{Z}^+$, but it does not contain $\mathbf{F}(\mathbf{x}_0)$. Thus, the set $D = \mathbf{F}^{-1}(C)$ contains the sequence $\{\mathbf{x}_k\}$, but does not contain its limit $\mathbf{x}_0$. This means D is not closed. Therefore, (b) does not hold.

There is a much easier proof of Theorem 2.27 if $\mathcal{A} = \mathbb{R}^n$, using Theorem 2.26, and the fact that a set is closed if and only if its complement is open.

Theorem 2.26 and Theorem 2.27 provide useful tools to justfy that a set is open or closed in $\mathbb{R}^n$, using our known library of continuous functions.

Example 2.26

Let A be the subset of $\mathbb{R}^2$ given by

$$A = \left\{(x, y) \,|\, x^2 + y^2 < 20, y > x^2\right\}.$$

Show that A is open.

Solution

Let $\mathcal{O} = \{(x, y) \,|\, x^2 + y^2 < 20\}$. This is a ball of radius $\sqrt{20}$ centered at the origin. Hence, $\mathcal{O}$ is open. Define the function $f : \mathcal{O} \to \mathbb{R}$ by $f(x, y) = y - x^2$. Since f is a polynomial, it is continuous. Notice that $y > x^2$ if and only if $f(x, y) > 0$, if and only if $f(x, y) \in (0, \infty)$. This shows that $A = f^{-1}((0, \infty))$. Since $(0, \infty)$ is open in $\mathbb{R}$, Theorem 2.26 implies that A is an open set.

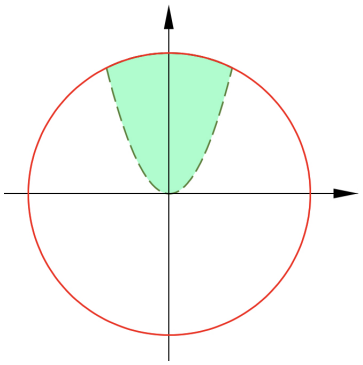


Figure 2.12: The set A in Example 2.26.

Example 2.27

Let C be the subset of $\mathbb{R}^3$ given by

$$C = \left\{(x, y, z) \,|\, x \geq 0, y \geq 0, y^2 + z^2 \leq 20.\right\}.$$

Show that C is closed.

Solution

Let $\pi_x : \mathbb{R}^3 \to \mathbb{R}$ and $\pi_y : \mathbb{R}^3 \to \mathbb{R}$ be the projection functions $\pi_x(x, y, z) = x$ and $\pi_y(x, y, z) = y$, and consider the function $g : \mathbb{R}^3 \to \mathbb{R}$ defined as

$$g(x, y, z) = 20 - (y^2 + z^2).$$

Notice that $y^2 + z^2 \leq 20$ if and only if $g(x, y, z) \geq 0$, if and only if $g(x, y, z) \in I = [0, \infty)$. The projection functions π_x and π_y are continuous. Since g is a polynomial, it is also continuous. The set $I = [0, \infty)$ is closed in $\mathbb{R}$. Therefore, the sets $\pi_x^{-1}(I)$, $\pi_y^{-1}(I)$ and $g^{-1}(I)$ are closed in $\mathbb{R}^3$. Since

$$A = \pi_x^{-1}(I) \cap \pi_y^{-1}(I) \cap g^{-1}(I),$$

being an intersection of three closed sets, A is closed in $\mathbb{R}^3$.

Using the same reasonings, we obtain the following.

Theorem 2.28

Let $I_1, \ldots, I_n$ be intervals in $\mathbb{R}$.

1. If each of $I_1, \ldots, I_n$ are open intervals of the form (a, b), (a, ∞), $(-\infty, a)$ or $\mathbb{R}$, then $I_1 \times \cdots \times I_n$ is an open subset of $\mathbb{R}^n$.
2. If each of $I_1, \ldots, I_n$ are closed intervals of the form $[a, b]$, $[a, \infty)$, $(-\infty, a]$ or $\mathbb{R}$, then $I_1 \times \cdots \times I_n$ is a closed subset of $\mathbb{R}^n$.

Sketch of Proof

Use the fact that

$$I_1 \times \cdots \times I_n = \bigcap_{i=1}^{n} \pi_i^{-1}(I_i),$$

where $\pi_i : \mathbb{R}^n \to \mathbb{R}$ is the projection function $\pi_i(x_1, \ldots, x_n) = x_i$.

Example 2.28

The set

$$A = \{(x, y, z) \,|\, x < 0, y > 2, -10 < z < -3\}$$

is open in $\mathbb{R}^3$, since

$$A = (-\infty, 0) \times (2, \infty) \times (-10, -3).$$

The set

$$C = \{(x, y, z) \,|\, x \leq 0, y \geq 2, -10 \leq z \leq -3\}$$

is closed in $\mathbb{R}^3$, since

$$C = (-\infty, 0] \times [2, \infty) \times [-10, -3].$$

We also have the following.

Theorem 2.29

Let a and b be real numbers, and assume that $f : \mathbb{R}^n \to \mathbb{R}$ is a continuous function. Define the sets A, B, C, D, E and F as follows.

(a) $A = \{\mathbf{x} \in \mathbb{R}^n \,|\, f(\mathbf{x}) > a\}$

(b) $B = \{\mathbf{x} \in \mathbb{R}^n \,|\, f(\mathbf{x}) \geq a\}$

(c) $C = \{\mathbf{x} \in \mathbb{R}^n \,|\, f(\mathbf{x}) < a\}$

(d) $D = \{\mathbf{x} \in \mathbb{R}^n \,|\, f(\mathbf{x}) \leq a\}$

(e) $E = \{\mathbf{x} \in \mathbb{R}^n \,|\, a < f(\mathbf{x}) < b\}$

(f) $F = \{\mathbf{x} \in \mathbb{R}^n \,|\, a \leq f(\mathbf{x}) \leq b\}$

Then A, C and E are open sets, while B, D and F are closed sets.

The proof is left as an exercise.

Example 2.29

Find the interior, exterior and boundary of each of the following sets.

(a) $A = \{(x, y) \,|\, 0 < x^2 + 4y^2 < 4\}$

(b) $B = \{(x, y) \,|\, 0 < x^2 + 4y^2 \leq 4\}$

(c) $C = \{(x, y) \,|\, x^2 + 4y^2 \leq 4\}$

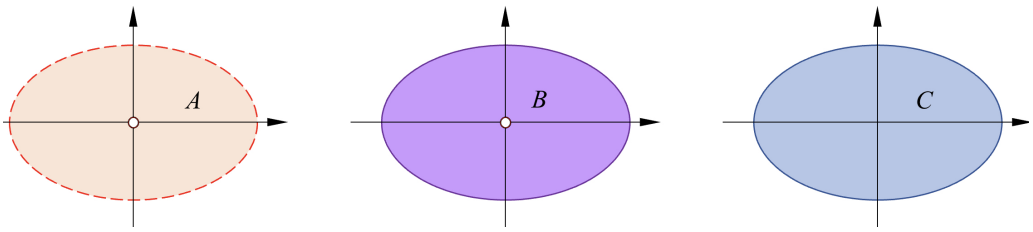


Figure 2.13: The sets A, B and C defined in Example 2.29.

Solution

Let

$$D = \left\{(x, y) \,|\, x^2 + 4y^2 < 4\right\}, \qquad E = \left\{(x, y) \,|\, x^2 + 4y^2 > 4\right\},$$

and let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function defined as

$$f(x, y) = x^2 + 4y^2.$$

Since f is a polynomial, it is continuous. By Theorem 2.29, A, D and E are open sets and C is a closed set. Since $A \subset B$ and $D \subset C$, we have

$$A = \operatorname{int} A \subset \operatorname{int} B \subset B, \qquad D \subset \operatorname{int} C.$$

Since $E = \mathbb{R}^2 \setminus C \subset \mathbb{R}^2 \setminus B \subset \mathbb{R}^2 \setminus A$, We have

$$E = \operatorname{ext} C \subset \operatorname{ext} B \subset \operatorname{ext} A.$$

Let

$$F = \left\{(x, y) \,|\, x^2 + 4y^2 = 4\right\}.$$

Then $\mathbb{R}^n$ is a disjoint union of D, E and F. If $\mathbf{u}_0 = (x_0, y_0) \in F$, either $x_0 \neq 0$ or $y_0 \neq 0$, but not both. If $x_0 \neq 0$, define the sequences $\{\mathbf{u}_k\}$ and $\{\mathbf{v}_k\}$ by

$$\mathbf{u}_k = \left(\frac{k}{k+1}x_0, y_0\right), \qquad \mathbf{v}_k = \left(\frac{k+1}{k}x_0, y_0\right).$$

If $x_0 = 0$, then $y_0 \neq 0$. Define the sequences $\{\mathbf{u}_k\}$ and $\{\mathbf{v}_k\}$ by

$$\mathbf{u}_k = \left(x_0, \frac{k}{k+1}y_0\right), \qquad \mathbf{v}_k = \left(x_0, \frac{k+1}{k}y_0\right).$$

In either case, $\{\mathbf{u}_k\}$ is a sequence of points in A that converges to $\mathbf{u}_0$, while $\{\mathbf{v}_k\}$ is a sequence of points in E that converges to $\mathbf{u}_0$. This proves that $\mathbf{u}_0$ is a boundary point of A, B and C. For the point $\mathbf{0}$, since it is not in A and B, it is not an interior point of A and B, but it is the limit of the sequence $\{(1/k, 0)\}$ that is in both A and B. Hence, $\mathbf{0}$ is in the closure of A and B, and hence, is a boundary point of A and B. We conclude that

$$\operatorname{int} A = \operatorname{int} B = \left\{(x,y) \,|\, 0 < x^2 + 4y^2 < 4\right\},$$
$$\operatorname{int} C = \left\{(x,y) \,|\, x^2 + 4y^2 < 4\right\},$$
$$\operatorname{ext} A = \operatorname{ext} B = \operatorname{ext} C = \left\{(x,y) \,|\, x^2 + 4y^2 > 4\right\},$$
$$\operatorname{bd} A = \operatorname{bd} B = \left\{(x,y) \,|\, x^2 + 4y^2 = 4\right\} \cup \{\mathbf{0}\},$$
$$\operatorname{bd} C = \left\{(x,y) \,|\, x^2 + 4y^2 = 4\right\}.$$

Remark 2.1

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function and let

$$C = \{\mathbf{x} \in \mathbb{R}^n \,|\, a \le f(\mathbf{x}) \le b\}.$$

One is tempting to say that

$$\operatorname{bd} C = \{\mathbf{x} \in \mathbb{R}^n \,|\, f(\mathbf{x}) = a \text{ or } f(\mathbf{x}) = b\}.$$

This is not necessary true. For example, consider the set C in Example 2.29. It can be written as

$$C = \left\{(x,y) \,|\, 0 \le x^2 + 4y^2 \le 4\right\}$$

However, the point where $f(x,y) = x^2 + 4y^2 = 0$ is not a boundary point of C.

Now we return to continuous functions.

Theorem 2.30 Pasting of Continuous Functions

Let A and B be closed subsets of $\mathbb{R}^n$, and let $S = A \cup B$. If $\mathbf{F} : S \to \mathbb{R}^m$ is a function such that $\mathbf{F}_A = \mathbf{F}|_A : A \to \mathbb{R}^m$ and $\mathbf{F}_B = \mathbf{F}|_B : B \to \mathbb{R}^m$ are both continuous, then $\mathbf{F} : S \to \mathbb{R}^m$ is continuous.

Proof

Since S is a union of two closed sets, it is closed. Applying Theorem 2.27, it suffices to show that if C is a closed subset of $\mathbb{R}^m$, then $\mathbf{F}^{-1}(C)$ is closed in $\mathbb{R}^n$. Notice that

$$\begin{aligned}\mathbf{F}^{-1}(C) &= \{\mathbf{x} \in S \,|\, \mathbf{F}(\mathbf{x}) \in C\} \\ &= \{\mathbf{x} \in A \,|\, \mathbf{F}(\mathbf{x}) \in C\} \cup \{\mathbf{x} \in B \,|\, \mathbf{F}(\mathbf{x}) \in C\} \\ &= \mathbf{F}_A^{-1}(C) \cup \mathbf{F}_B^{-1}(C).\end{aligned}$$

Since $\mathbf{F}_A : A \to \mathbb{R}^m$ and $\mathbf{F}_B : B \to \mathbb{R}^m$ are both continuous functions, $\mathbf{F}_A^{-1}(C)$ and $\mathbf{F}_B^{-1}(C)$ are closed subsets of $\mathbb{R}^n$. Being a union of two closed subsets, $\mathbf{F}^{-1}(C)$ is closed. This completes the proof.

Example 2.30

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function defined as

$$f(x,y) = \begin{cases} x^2 + y^2, & \text{if } x^2 + y^2 < 1 \\ 1, & \text{if } x^2 + y^2 \geq 1. \end{cases}$$

Show that f is a continuous function.

Solution

Let $A = \{(x,y) \,|\, x^2 + y^2 \leq 1\}$ and $B = \{(x,y) \,|\, x^2 + y^2 \geq 1\}$. Then A and B are closed subsets of $\mathbb{R}^2$ and $\mathbb{R}^2 = A \cup B$. Notice that $f|_A : A \to \mathbb{R}$ is the function $f(x,y) = x^2 + y^2$, which is continuous since it is a polynomial. By definition, $f|_B : B \to \mathbb{R}$ is the constant function $f_B(x,y) = 1$, which is also continuous. By Theorem 2.30, the function $f : \mathbb{R}^2 \to \mathbb{R}$ is continuous.

Given positive integers n and m, there is a natural bijective correspondence between $\mathbb{R}^n \times \mathbb{R}^m$ and $\mathbb{R}^{n+m}$ given by $\mathbf{T} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{n+m}$,

$$(\mathbf{x}, \mathbf{y}) \mapsto (x_1, \ldots, x_n, y_1, \ldots, y_m),$$

where

$$\mathbf{x} = (x_1, \ldots, x_n) \quad \text{and} \quad \mathbf{y} = (y_1, \ldots, y_m).$$

Hence, sometimes we will denote a point in $\mathbb{R}^{n+m}$ as $(\mathbf{x}, \mathbf{y})$, where $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$. By generalized Pythagoras theorem,

$$\|(\mathbf{x}, \mathbf{y})\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2.$$

If A is a subset of $\mathbb{R}^n$, B is a subset of $\mathbb{R}^m$, $A \times B$ can be considered as a subset of $\mathbb{R}^{n+m}$ given by

$$A \times B = \{(\mathbf{x}, \mathbf{y}) \,|\, \mathbf{x} \in A, \mathbf{y} \in B\}.$$

The following is more general than Proposition 2.16.

Proposition 2.31

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$, and let $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^k$ and $\mathbf{G} : \mathfrak{D} \to \mathbb{R}^l$ be functions defined on $\mathfrak{D}$. Define the function $\mathbf{H} : \mathfrak{D} \to \mathbb{R}^{k+l}$ by

$$\mathbf{H}(\mathbf{x}) = (\mathbf{F}(\mathbf{x}), \mathbf{G}(\mathbf{x})).$$

Then the function $\mathbf{H} : \mathfrak{D} \to \mathbb{R}^{k+l}$ is continuous if and only if the functions $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^k$ and $\mathbf{G} : \mathfrak{D} \to \mathbb{R}^l$ are continuous.

Sketch of Proof

This proposition follows immediately from Proposition 2.16, since

$$\mathbf{H}(\mathbf{x}) = (F_1(\mathbf{x}), \ldots, F_k(\mathbf{x}), G_1(\mathbf{x}), \ldots, G_l(\mathbf{x})).$$

For a function defined on a subset of $\mathbb{R}^n$, we can define its graph in the following way.

Definition 2.8 The Graph of a Function

Let $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ be a function defined on $\mathfrak{D} \subset \mathbb{R}^n$. The graph of $\mathbf{F}$, denoted by $G_{\mathbf{F}}$, is the subset of $\mathbb{R}^{n+m}$ defined as

$$G_{\mathbf{F}} = \{(\mathbf{x}, \mathbf{y}) \,|\, \mathbf{x} \in \mathfrak{D}, \mathbf{y} = \mathbf{F}(\mathbf{x})\}.$$

Example 2.31

Let $\mathfrak{D} = \{(x, y) \,|\, x^2 + y^2 \leq 1\}$, and let $f : \mathfrak{D} \to \mathbb{R}$ be the function defined as

$$f(x, y) = \sqrt{1 - x^2 - y^2}.$$

The graph of f is

$$G_f = \left\{(x, y, z) \,|\, x^2 + y^2 \leq 1, z = \sqrt{1 - x^2 - y^2}\right\},$$

which is the upper hemisphere.

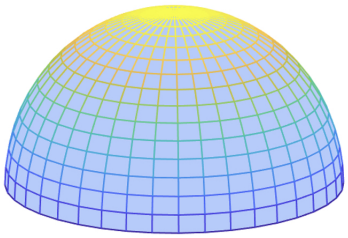


Figure 2.14: The upper hemisphere is the graph of a function.

Notice that if $\mathfrak{D}$ is a subset of $\mathbb{R}^n$, then the graph of the function $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is the image of the function $\mathbf{H} : \mathfrak{D} \to \mathbb{R}^{n+m}$ defined as

$$\mathbf{H}(\mathbf{x}) = (\mathbf{x}, \mathbf{F}(\mathbf{x})).$$

From Proposition 2.31, we obtain the following.

Corollary 2.32

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$, and let $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ be a function defined on $\mathfrak{D}$. The image of the function $\mathbf{H} : \mathfrak{D} \to \mathbb{R}^{n+m}$,

$$\mathbf{H}(\mathbf{x}) = (\mathbf{x}, \mathbf{F}(\mathbf{x})),$$

is the graph of $\mathbf{F}$. If the function $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is continuous, then the function $\mathbf{H} : \mathfrak{D} \to \mathbb{R}^{n+m}$ is continuous.

Now we consider a special class of functions called Lipschitz functions.

Definition 2.9

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$. A function $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is Lipschitz provided that there exists a positive constant c such that

$$\|\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v})\| \leq c\|\mathbf{u} - \mathbf{v}\| \qquad \text{for all } \mathbf{u}, \mathbf{v} \in \mathfrak{D}.$$

The constant c is called a Lipschitz constant of the function. If $c < 1$, then $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is called a contraction.

The following is easy to establish.

Proposition 2.33

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$, and let $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ be a Lipschitz function. Then $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is continuous.

Example 2.32

A linear transformation of the form $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{T}(\mathbf{x}) = a\mathbf{x}$, is a Lipschitz function with Lipschitz constant $|a|$.

In fact, we have the following.

Theorem 2.34

A linear transformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^m$ is a Lipschitz function.

Proof

Let A be the $m \times n$ matrix such that $\mathbf{T}(\mathbf{x}) = A\mathbf{x}$. When $\mathbf{x}$ is in $\mathbb{R}^n$,

$$\|\mathbf{T}(\mathbf{x})\|^2 = (A\mathbf{x})^T (A\mathbf{x}) = \mathbf{x}^T (A^T A)\mathbf{x}.$$

The matrix $B = A^T A$ is a positive semi-definite $n \times n$ symmetric matrix. By Theorem 2.7,

$$\mathbf{x}^T (A^T A)\mathbf{x} \leq \lambda_{\max} \|\mathbf{x}\|^2,$$

where $\lambda_{\max}$ is the largest eigenvalue of $A^T A$.

Therefore, for any $\mathbf{x} \in \mathbb{R}^n$,

$$\|\mathbf{T}(\mathbf{x})\| \leq \sqrt{\lambda_{\max}}\|\mathbf{x}\|.$$

It follows that for any $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$,

$$\|\mathbf{T}(\mathbf{u}) - \mathbf{T}(\mathbf{v})\| = \|\mathbf{T}(\mathbf{u} - \mathbf{v})\| \leq \sqrt{\lambda_{\max}}\|\mathbf{u} - \mathbf{v}\|.$$

Hence, $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^m$ is a Lipschitz mapping with Lipschitz constant $\sqrt{\lambda_{\max}}$.

Example 2.33

Let $\mathbf{T} : \mathbb{R}^2 \to \mathbb{R}^2$ be the mapping defined as

$$\mathbf{T}(x, y) = (x - 3y, 7x + 4y).$$

Find the smallest constant c such that

$$\|\mathbf{T}(\mathbf{u}) - \mathbf{T}(\mathbf{v})\| \leq c\|\mathbf{u} - \mathbf{v}\|$$

for all $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^2$.

Solution

Notice that $\mathbf{T}(\mathbf{u}) = A\mathbf{u}$, where A is the 2×2 matrix $A = \begin{bmatrix} 1 & -3 \\ 7 & 4 \end{bmatrix}$. Hence,

$$\|\mathbf{T}(\mathbf{u})\|^2 = \mathbf{u}^T A^T A\mathbf{u} = \mathbf{u}^T C\mathbf{u},$$

where

$$C = \begin{bmatrix} 1 & 7 \\ -3 & 4 \end{bmatrix} \begin{bmatrix} 1 & -3 \\ 7 & 4 \end{bmatrix} = \begin{bmatrix} 50 & 25 \\ 25 & 25 \end{bmatrix} = 25 \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}.$$

For the matrix $G = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$, the eigenvalues are the solutions of

$$\lambda^2 - 3\lambda + 1 = 0,$$

which are

$$\lambda_1 = \frac{3+\sqrt{5}}{2} \quad \text{and} \quad \lambda_2 = \frac{3-\sqrt{5}}{2}.$$

Hence,

$$\|\mathbf{T}(\mathbf{u})\|^2 \leq \frac{25(3+\sqrt{5})}{2}\|\mathbf{u}\|^2.$$

The smallest c such that $\|\mathbf{T}(\mathbf{u}) - \mathbf{T}(\mathbf{v})\| \leq c\|\mathbf{u} - \mathbf{v}\|$ for all $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^2$ is

$$c = \sqrt{\frac{25(3+\sqrt{5})}{2}} = 8.0902.$$

Remark 2.2

If A is an $m \times n$ matrix, the matrix $B = A^T A$ is a positive semi-definite $n \times n$ symmetric matrix. Thus, all its eigenvalues are nonnegative. Let $\lambda_1, \ldots, \lambda_n$ be its eigenvalues with

$$0 = \lambda_n = \cdots = \lambda_{r+1} < \lambda_r \leq \lambda_{r-1} \leq \cdots \leq \lambda_1.$$

Then $\lambda_1, \cdots, \lambda_r$ are the nonzero eigenvalues of $A^T A$. The singular values of A are the numbers $\sigma_1, \ldots, \sigma_r$, where

$$\sigma_i = \sqrt{\lambda_i}, \quad 1 \leq i \leq r.$$

Theorem 2.34 says that σ_1 is a Lipschitz constant of the linear transformation $\mathbf{T}(\mathbf{x}) = A\mathbf{x}$.

At the end of this section, we want to discuss the vector space of $m \times n$ matrices $\mathcal{M}_{m,n}$. There is a natural vector space isomorphism between $\mathcal{M}_{m,n}$ and $\mathbb{R}^{mn}$, by mapping the matrix $A = [a_{ij}]$ to $\mathbf{x} = (x_k)$, where

$$x_{(i-1)n+j} = a_{ij} \qquad \text{for } 1 \leq i \leq m, 1 \leq j \leq n.$$

In other words, if

$$\begin{aligned}\mathbf{a}_1 &= (a_{11}, a_{12}, \ldots, a_{1n}),\\ \mathbf{a}_2 &= (a_{21}, a_{22}, \ldots, a_{2n}),\\ &\vdots\\ \mathbf{a}_m &= (a_{m,1}, a_{m,2}, \ldots, a_{m,n})\end{aligned}$$

are the row vectors of A, then A is mapped to the vector $(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m)$ in $\mathbb{R}^{mn}$. Under this isomorphism, the norm of a matrix $A = [a_{ij}]$ is

$$\|A\| = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2} = \sqrt{\sum_{i=1}^{m} \|\mathbf{a}_i\|^2},$$

and the distance between two matrices $A = [a_{ij}]$ and $B = [b_{ij}]$ is

$$d(A, B) = \|A - B\| = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} (a_{ij} - b_{ij})^2}.$$

The following proposition can be used to give an alternative proof of Theorem 2.34.

Proposition 2.35

Let A be an $m \times n$ matrix. If $\mathbf{x}$ is in $\mathbb{R}^n$, then

$$\|A\mathbf{x}\| \leq \|A\|\|\mathbf{x}\|.$$

Proof

Let $\mathbf{a}_1, \ldots, \mathbf{a}_m$ be the row vectors of A, and let $\mathbf{w} = A\mathbf{x}$. Then

$$w_i = \langle \mathbf{a}_i, \mathbf{x} \rangle \qquad \text{for } 1 \leq i \leq m.$$

By Cauchy-Schwarz inequality,

$$|w_i| \leq \|\mathbf{a}_i\|\|\mathbf{x}\| \qquad \text{for } 1 \leq i \leq m.$$

Thus,

$$\begin{aligned}\|\mathbf{w}\| &= \sqrt{w_1^2 + w_2^2 + \cdots + w_m^2} \\ &\leq \|\mathbf{x}\| \sqrt{\|\mathbf{a}_1\|^2 + \|\mathbf{a}_2\|^2 + \cdots + \|\mathbf{a}_m\|^2} = \|A\|\|\mathbf{x}\|.\end{aligned}$$

The difference between the proofs of Theorem 2.34 and Proposition 2.35 is that, in the proof of Theorem 2.34, we find that the smallest possible c such that $\|A\mathbf{x}\| \leq c\|\mathbf{x}\|$ for all $\mathbf{x}$ in $\mathbb{R}^n$ is the largest singular value of the matrix A. In Proposition 2.35, we find a candidate for c, which is the norm of the matrix A, but this is usually not the optimal one.

When $m = n$, we denote the space of $n \times n$ matrices $\mathcal{M}_{n,n}$ simply as $\mathcal{M}_n$. The determinant of the matrix $A = [a_{ij}] \in \mathcal{M}_n$ is given by

$$\det A = \sum_{\sigma} \operatorname{sgn}(\sigma) a_{1\sigma(1)} a_{2\sigma(2)} \cdots a_{n\sigma(n)}.$$

Here the summation is over all the $n!$ permutations σ of the set $S_n = \{1, 2, \ldots, n\}$, and $\operatorname{sgn}(\sigma)$ is the sign of the permutation σ, which is equal to 1 or -1, depending on whether σ can be written as the product of an even number or an odd number of transpositions. For example, when $n = 1$, $\det[a] = a$. When $n = 2$,

$$\det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$

When $n = 3$,

$$\begin{aligned}\det \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} &= a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ &\quad - a_{11}a_{23}a_{32} - a_{13}a_{22}a_{31} - a_{12}a_{21}a_{33}.\end{aligned}$$

The determinant function $\det : \mathcal{M}_n \to \mathbb{R}$ is a polynomial function on the variables (a_{ij}). Hence, it is a continuous function. Recall that a matrix $A \in \mathcal{M}_n$ is invertible if and only if $\det A \neq 0$. Let

$$\mathrm{GL}\,(n, \mathbb{R}) = \{A \in \mathcal{M}_n \mid \det A \neq 0\}$$

be the subset of $\mathcal{M}_n$ that consist of invertible $n \times n$ matrices. It is a group under matrix multiplication, called the general linear group. By definition,

$$\mathrm{GL}\,(n, \mathbb{R}) = \det^{-1}(\mathbb{R} \setminus \{0\}).$$

Since $\mathbb{R} \setminus \{0\}$ is an open subset of $\mathbb{R}$, $\mathrm{GL}\,(n, \mathbb{R})$ is an open subset of $\mathcal{M}_n$. This gives the following.

Proposition 2.36

Given that A is an invertible $n \times n$ matrix, there exists $r > 0$ such that if B is an $n \times n$ matrix with $\|B - A\| < r$, then B is also invertible.

Sketch of Proof

This is simply a rephrase of the statement that if A is a point in the open set $\mathrm{GL}\,(n, \mathbb{R})$, then there is a ball $B(A, r)$ with center at A that is contained in $\mathrm{GL}\,(n, \mathbb{R})$.

Let A be an $n \times n$ matrix. For $1 \leq i, j \leq n$, the (i, j)-minor of A, denoted by $M_{i,j}$, is the determinant of the $(n-1) \times (n-1)$ matrix obtained by deleting the i^{th}-row and j^{th}- column of A. Using the same reasoning as above, we find that the function $M_{i,j} : \mathcal{M}_n \to \mathbb{R}$ is a continuous function. The (i, j) cofactor $C_{i,j}$ of A is given by $C_{i,j} = (-1)^{i+j} M_{i,j}$. The cofactor matrix of A is $C_A = [C_{ij}]$. Since each of the components is continuous, the function $C : \mathcal{M}_n \to \mathcal{M}_n$ taking A to C_A is a continuous function.

If A is invertible,

$$A^{-1} = \frac{1}{\det A} C_A^T.$$

Since both $C : \mathcal{M}_n \to \mathcal{M}_n$ and $\det : \mathcal{M}_n \to \mathbb{R}$ are continuous functions, and $\det : \mathrm{GL}\,(n, \mathbb{R}) \to \mathbb{R}$ is a function that is never equal to 0, we obtain the following.

Theorem 2.37

The map $\mathscr{I} : \mathrm{GL}\,(n, \mathbb{R}) \to \mathrm{GL}\,(n, \mathbb{R})$ that takes A to A^{-1} is continuous.

Exercises 2.3

Question 1

Let $\mathbf{x}_0$ be a point in $\mathbb{R}^n$. Define the function $f : \mathbb{R}^n \to \mathbb{R}$ by

$$f(\mathbf{x}) = \|\mathbf{x} - \mathbf{x}_0\|.$$

Show that f is a continuous function.

Question 2

Let $\mathcal{O} = \mathbb{R}^3 \setminus \{(0,0,0)\}$ and define the function $\mathbf{F} : \mathcal{O} \to \mathbb{R}^2$ by

$$\mathbf{F}(x,y,z) = \left(\frac{y}{x^2+y^2+z^2}, \frac{z}{x^2+y^2+z^2} \right).$$

Show that $\mathbf{F}$ is a continuous function.

Question 3

Let $f : \mathbb{R}^n \to \mathbb{R}$ be the function defined as

$$f(\mathbf{x}) = \begin{cases} 1, & \text{if at least one of the } x_i \text{ is rational,} \\ 0, & \text{otherwise.} \end{cases}$$

At which point of $\mathbb{R}^n$ is the function f continuous?

Question 4

Let $f : \mathbb{R}^n \to \mathbb{R}$ be the function defined as

$$f(\mathbf{x}) = \begin{cases} x_1^2 + \cdots + x_n^2, & \text{if at least one of the } x_i \text{ is rational,} \\ 0, & \text{otherwise.} \end{cases}$$

At which point of $\mathbb{R}^n$ is the function f continuous?

Question 5

Let $f : \mathbb{R}^3 \to \mathbb{R}$ be the function defined by

$$f(x,y,z) = \begin{cases} \dfrac{\sin(x^2+4y^2+z^2)}{x^2+4y^2+z^2}, & \text{if } (x,y,z) \neq (0,0,0), \\ a, & \text{if } (x,y,z) = (0,0,0). \end{cases}$$

Show that there exists a value a such that f is a continuous function, and find this value of a.

Question 6

Let a and b be positive numbers, and let $\mathcal{O}$ be the subset of $\mathbb{R}^n$ defined as

$$\mathcal{O} = \{\mathbf{x} \in \mathbb{R}^n \,|\, a < \|\mathbf{x}\| < b\}\,.$$

Show that $\mathcal{O}$ is open.

Question 7

Let A be the subset of $\mathbb{R}^2$ given by

$$A = \{(x,y) \,|\, \sin(x+y) + xy > 1\}\,.$$

Show that A is an open set.

Question 8

Let A be the subset of $\mathbb{R}^3$ given by

$$A = \{(x,y,z) \,|\, x \geq 0, y \leq 1, e^{xy} \leq z\}\,.$$

Show that A is a closed set.

Question 9

A plane in $\mathbb{R}^3$ is the set of all points (x, y, z) satisfying an equation of the form

$$ax + by + cz = d,$$

where $(a, b, c) \neq (0, 0, 0)$. Show that a plane is a closed subset of $\mathbb{R}^3$.

Question 10

Define the sets A, B, C and D as follows.

(a) $A = \{(x, y, z) \,|\, x^2 + 4y^2 + 9z^2 < 36\}$

(b) $B = \{(x, y, z) \,|\, x^2 + 4y^2 + 9z^2 \leq 36\}$

(c) $C = \{(x, y, z) \,|\, 0 < x^2 + 4y^2 + 9z^2 < 36\}$

(d) $D = \{(x, y, z) \,|\, 0 < x^2 + 4y^2 + 9z^2 \leq 36\}$

For each of these sets, find its interior, exterior and boundary.

Question 11

Let a and b be real numbers, and assume that $f : \mathbb{R}^n \to \mathbb{R}$ is a continuous function. Consider the following subsets of $\mathbb{R}^n$.

(a) $A = \{\mathbf{x} \in \mathbb{R}^n \,|\, f(\mathbf{x}) > a\}$

(b) $B = \{\mathbf{x} \in \mathbb{R}^n \,|\, f(\mathbf{x}) \geq a\}$

(c) $C = \{\mathbf{x} \in \mathbb{R}^n \,|\, f(\mathbf{x}) < a\}$

(d) $D = \{\mathbf{x} \in \mathbb{R}^n \,|\, f(\mathbf{x}) \leq a\}$

(e) $E = \{\mathbf{x} \in \mathbb{R}^n \,|\, a < f(\mathbf{x}) < b\}$

(f) $F = \{\mathbf{x} \in \mathbb{R}^n \,|\, a \leq f(\mathbf{x}) \leq b\}$

Show that A, C and E are open sets, while B, D and F are closed sets.

Question 12

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function defined as

$$f(x,y) = \begin{cases} x^2 + y^2, & \text{if } x^2 + y^2 < 4 \\ 8 - x^2 - y^2, & \text{if } x^2 + y^2 \geq 4. \end{cases}$$

Show that f is a continuous function.

Question 13

Show that the distance function on $\mathbb{R}^n$, $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$,

$$d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|,$$

is continuous in the following sense. If $\{\mathbf{u}_k\}$ and $\{\mathbf{v}_k\}$ are sequences in $\mathbb{R}^n$ that converges to $\mathbf{u}$ and $\mathbf{v}$ respectively, then the sequence $\{d(\mathbf{u}_k, \mathbf{v}_k)\}$ converges to $d(\mathbf{u}, \mathbf{v})$.

Question 14

Let $\mathbf{T} : \mathbb{R}^2 \to \mathbb{R}^3$ be the mapping

$$\mathbf{T}(x,y) = (x + y, 3x - y, 6x + 5y).$$

Show that $\mathbf{T} : \mathbb{R}^2 \to \mathbb{R}^3$ is a Lipschitz mapping, and find the smallest Lipschitz constant for this mapping.

Question 15

Given that A is a subset of $\mathbb{R}^m$ and B is a subset of $\mathbb{R}^n$, let $C = A \times B$. Then C is a subset of $\mathbb{R}^{m+n}$.

(a) If A is open in $\mathbb{R}^m$ and B is open in $\mathbb{R}^n$, show that $A \times B$ is open in $\mathbb{R}^{m+n}$.

(b) If A is closed in $\mathbb{R}^m$ and B is closed in $\mathbb{R}^n$, show that $A \times B$ is closed in $\mathbb{R}^{m+n}$.

Question 16

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$, and let $f : \mathfrak{D} \to \mathbb{R}$ be a continuous function defined on $\mathfrak{D}$. Let $A = \mathfrak{D} \times \mathbb{R}$ and define the function $g : A \to \mathbb{R}$ by

$$g(\mathbf{x}, y) = y - f(\mathbf{x}).$$

Show that $g : A \to \mathbb{R}$ is continuous.

Question 17

Let U be an open subset of $\mathbb{R}^n$, and let $f : U \to \mathbb{R}$ be a continuous function defined on U. Show that the sets

$$\mathcal{O}_1 = \{(\mathbf{x}, y) \,|\, \mathbf{x} \in U, y < f(\mathbf{x})\}, \quad \mathcal{O}_2 = \{(\mathbf{x}, y) \,|\, \mathbf{x} \in U, y > f(\mathbf{x})\}$$

are open subsets of $\mathbb{R}^{n+1}$.

Question 18

Let C be a closed subset of $\mathbb{R}^n$, and let $f : C \to \mathbb{R}$ be a continuous function defined on C. Show that the sets

$$\mathcal{A}_1 = \{(\mathbf{x}, y) \,|\, \mathbf{x} \in C, y \leq f(\mathbf{x})\}, \quad \mathcal{A}_2 = \{(\mathbf{x}, y) \,|\, \mathbf{x} \in C, y \geq f(\mathbf{x})\}$$

are closed subsets of $\mathbb{R}^{n+1}$.

2.4 Uniform Continuity

In volume I, we have seen that uniform continuity plays important role in single variable analysis. In this section, we extend this concept to multivariable functions.

Definition 2.10 Uniform Continuity

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$, and let $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ be a function defined on $\mathfrak{D}$. We say that the function $\mathbf{F}$ is **uniformly continuous** provided that for any $\varepsilon > 0$, there exists $\delta > 0$ such that if $\mathbf{u}$ and $\mathbf{v}$ are points in $\mathfrak{D}$ and $\|\mathbf{u} - \mathbf{v}\| < \delta$, then

$$\|\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v})\| < \varepsilon.$$

The following two propositions are obvious.

Proposition 2.38

A uniformly continuous function is continuous.

Proposition 2.39

Given that $\mathfrak{D}$ is a subset of $\mathbb{R}^n$, and $\mathfrak{D}'$ is a subset of $\mathfrak{D}$, if the function $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is uniformly continuous, then the function $\mathbf{F} : \mathfrak{D}' \to \mathbb{R}^m$ is also uniformly continuous.

A special class of uniformly continuous functions is the class of Lipschitz functions.

Theorem 2.40

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$, and let $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ be a function defined on $\mathfrak{D}$. If $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is Lipschitz, then it is uniformly continuous.

The proof is straightforward.

Remark 2.3

Theorem 2.34 and Theorem 2.40 imply that a linear transformation is uniformly continuous.

There is an equivalent definition for uniform continuity in terms of sequences.

Theorem 2.41

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$, and let $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ be a function defined on $\mathfrak{D}$. Then the following are equivalent.

(i) $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is uniformly continuous. Namely, given $\varepsilon > 0$, there exists $\delta > 0$ such that if $\mathbf{u}$ and $\mathbf{v}$ are points in $\mathfrak{D}$ and $\|\mathbf{u} - \mathbf{v}\| < \delta$, then

$$\|\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v})\| < \varepsilon.$$

(ii) If $\{\mathbf{u}_k\}$ and $\{\mathbf{v}_k\}$ are two sequences in $\mathfrak{D}$ such that

$$\lim_{k\to\infty} (\mathbf{u}_k - \mathbf{v}_k) = \mathbf{0},$$

then

$$\lim_{k\to\infty} (\mathbf{F}(\mathbf{u}_k) - \mathbf{F}(\mathbf{v}_k)) = \mathbf{0}.$$

Let us give a proof of this theorem here.

Proof

Assume that (i) holds, and $\{\mathbf{u}_k\}$ and $\{\mathbf{v}_k\}$ are two sequences in $\mathfrak{D}$ such that

$$\lim_{k\to\infty} (\mathbf{u}_k - \mathbf{v}_k) = \mathbf{0}.$$

Given $\varepsilon > 0$, (i) implies that there exists $\delta > 0$ such that if $\mathbf{u}$ and $\mathbf{v}$ are points in $\mathfrak{D}$ and $\|\mathbf{u} - \mathbf{v}\| < \delta$, then

$$\|\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v})\| < \varepsilon.$$

Since $\lim_{k\to\infty} (\mathbf{u}_k - \mathbf{v}_k) = \mathbf{0}$, there is a positive integer K such that for all $k \geq K$, $\|\mathbf{u}_k - \mathbf{v}_k\| < \delta$. It follows that

$$\|\mathbf{F}(\mathbf{u}_k) - \mathbf{F}(\mathbf{v}_k)\| < \varepsilon \qquad \text{for all } k \geq K.$$

This shows that

$$\lim_{k\to\infty} (\mathbf{F}(\mathbf{u}_k) - \mathbf{F}(\mathbf{v}_k)) = \mathbf{0},$$

and thus completes the proof of (i) implies (ii).

Conversely, assume that (i) does not hold. This means there exists an $\varepsilon > 0$, for all $\delta > 0$, there exist points $\mathbf{u}$ and $\mathbf{v}$ in $\mathfrak{D}$ such that $\|\mathbf{u} - \mathbf{v}\| < \delta$ and $\|\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v})\| \geq \varepsilon$. Thus, for every $k \in \mathbb{Z}^+$, there exists $\mathbf{u}_k$ and $\mathbf{v}_k$ in $\mathfrak{D}$ such that

$$\|\mathbf{u}_k - \mathbf{v}_k\| < \frac{1}{k}, \tag{2.5}$$

and $\|\mathbf{F}(\mathbf{u}_k) - \mathbf{F}(\mathbf{v}_k)\| \geq \varepsilon$. Notice that $\{\mathbf{u}_k\}$ and $\{\mathbf{v}_k\}$ are sequences in $\mathfrak{D}$. Eq. (2.5) implies that $\lim_{k\to\infty} (\mathbf{u}_k - \mathbf{v}_k) = \mathbf{0}$. Since $\|\mathbf{F}(\mathbf{u}_k) - \mathbf{F}(\mathbf{v}_k)\| \geq \varepsilon$,

$$\lim_{k\to\infty} (\mathbf{F}(\mathbf{u}_k) - \mathbf{F}(\mathbf{v}_k)) \neq \mathbf{0}.$$

This shows that if (i) does not hold, then (ii) does not hold.

From Theorem 2.41, we can deduce the following.

Proposition 2.42

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$, and let $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ be a function defined on $\mathfrak{D}$. Then $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is uniformly continuous if and only if each of the component functions $F_j = (\pi_j \circ \mathbf{F}) : \mathfrak{D} \to \mathbb{R}$, $1 \leq j \leq m$, is uniformly continuous.

Let us look at some more examples.

Example 2.34

Let $\mathfrak{D}$ be the open rectangle $\mathfrak{D} = (0, 5) \times (0, 7)$, and consider the function $f : \mathfrak{D} \to \mathbb{R}$ defined by

$$f(x, y) = xy.$$

Determine whether $f : \mathfrak{D} \to \mathbb{R}$ is uniformly continuous.

Solution

For any two points $\mathbf{u}_1 = (x_1, y_1)$ and $\mathbf{u}_2 = (x_2, y_2)$ in $\mathfrak{D}$, $0 < x_1, x_2 < 5$ and $0 < y_1, y_2 < 7$. Since

$$f(\mathbf{u}_1) - f(\mathbf{u}_2) = x_1y_1 - x_2y_2 = x_1(y_1 - y_2) + y_2(x_1 - x_2),$$

we find that

$$\begin{aligned} |f(\mathbf{u}_1) - f(\mathbf{u}_2)| &\leq |x_1||y_1 - y_2| + |y_2||x_1 - x_2| \\ &\leq 5\|\mathbf{u}_1 - \mathbf{u}_2\| + 7\|\mathbf{u}_1 - \mathbf{u}_2\| = 12\|\mathbf{u}_1 - \mathbf{u}_2\|. \end{aligned}$$

This shows that $f : \mathfrak{D} \to \mathbb{R}$ is a Lipschitz function. Hence, it is uniformly continuous.

Example 2.35

Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x, y) = xy.$$

Determine whether $f : \mathbb{R}^2 \to \mathbb{R}$ is uniformly continuous.

Solution

For $k \in \mathbb{Z}^+$, let

$$\mathbf{u}_k = \left(k + \frac{1}{k}, k\right), \qquad \mathbf{v}_k = (k, k).$$

Then $\{\mathbf{u}_k\}$ and $\{\mathbf{v}_k\}$ are sequences of points in $\mathbb{R}^2$ and

$$\lim_{k\to\infty} (\mathbf{u}_k - \mathbf{v}_k) = \lim_{k\to\infty} \left(\frac{1}{k}, 0\right) = (0, 0).$$

However,

$$f(\mathbf{u}_k) - f(\mathbf{v}_k) = k\left(k + \frac{1}{k}\right) - k^2 = 1.$$

Thus,

$$\lim_{k\to\infty} (f(\mathbf{u}_k) - f(\mathbf{v}_k)) = 1 \neq 0.$$

Therefore, the function $f : \mathbb{R}^2 \to \mathbb{R}$ is not uniformly continuous.

Example 2.34 and 2.35 show that whether a function is uniformly continuous depends on the domain of the function.

Exercises 2.4

Question 1

Let $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^2$ be the function defined as

$$\mathbf{F}(x, y, z) = (3x - 2z + 7, x + y + z - 4).$$

Show that $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^2$ is uniformly continuous.

Question 2

Let $\mathfrak{D} = (0, 1) \times (0, 2)$. Consider the function $f : \mathfrak{D} \to \mathbb{R}$ defined as

$$f(x, y) = x^2 + 3y.$$

Determine whether f is uniformly continuous.

Question 3

Let $\mathfrak{D} = (1, \infty) \times (1, \infty)$. Consider the function $f : \mathfrak{D} \to \mathbb{R}$ defined as

$$f(x, y) = \sqrt{x + y}.$$

Determine whether f is uniformly continuous.

Question 4

Let $\mathfrak{D} = (0, 1) \times (0, 2)$. Consider the function $f : \mathfrak{D} \to \mathbb{R}$ defined as

$$f(x, y) = \frac{1}{\sqrt{x + y}}.$$

Determine whether f is uniformly continuous.

2.5 Contraction Mapping Theorem

Among the Lipschitz functions, there is a subset called contractions.

Definition 2.11 Contractions

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$. A function $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is called a contraction if there exists a constant $0 \leq c < 1$ such that

$$\|\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v})\| \leq c\|\mathbf{u} - \mathbf{v}\| \qquad \text{for all } \mathbf{u}, \mathbf{v} \in \mathfrak{D}.$$

In other words, a contraction is a Lipschitz function which has a Lipschitz constant that is less than 1.

Example 2.36

Let $\mathbf{b}$ be a point in $\mathbb{R}^n$, and let $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^n$ be the function defined as

$$\mathbf{F}(\mathbf{x}) = c\mathbf{x} + \mathbf{b}.$$

The mapping $\mathbf{F}$ is a contraction if and only if $|c| < 1$.

The contraction mapping theorem is an important result in analysis. Extended to metric spaces, it is an important tool to prove the existence and uniqueness of solutions of ordinary differential equations.

Theorem 2.43 Contraction Mapping Theorem

Let $\mathfrak{D}$ be a closed subset of $\mathbb{R}^n$, and let $\mathbf{F} : \mathfrak{D} \to \mathfrak{D}$ be a contraction. Then $\mathbf{F}$ has a unique fixed point. Namely, there is a unique $\mathbf{u}$ in $\mathfrak{D}$ such that $\mathbf{F}(\mathbf{u}) = \mathbf{u}$.

Proof

By definition, there is a constant $c \in [0, 1)$ such that

$$\|\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v})\| \leq c\|\mathbf{u} - \mathbf{v}\| \qquad \text{for all } \mathbf{u}, \mathbf{v} \in \mathfrak{D}.$$

We start with any point $\mathbf{x}_0$ in $\mathfrak{D}$ and construct the sequence $\{\mathbf{x}_k\}$ inductively by

$$\mathbf{x}_{k+1} = \mathbf{F}(\mathbf{x}_k) \qquad \text{for all } k \geq 0.$$

Notice that for all $k \in \mathbb{Z}^+$,

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\| = \|\mathbf{F}(\mathbf{x}_k) - \mathbf{F}(\mathbf{x}_{k-1})\| \leq c\|\mathbf{x}_k - \mathbf{x}_{k-1}\|.$$

By iterating, we find that

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leq c^k\|\mathbf{x}_1 - \mathbf{x}_0\|.$$

Therefore, if $l > k \geq 0$, triangle inequality implies that

$$\begin{aligned}\|\mathbf{x}_l - \mathbf{x}_k\| &\leq \|\mathbf{x}_l - \mathbf{x}_{l-1}\| + \cdots + \|\mathbf{x}_{k+1} - \mathbf{x}_k\| \\ &\leq (c^{l-1} + \ldots + c^k)\|\mathbf{x}_1 - \mathbf{x}_0\|.\end{aligned}$$

Since $c \in [0, 1)$,

$$c^{l-1} + \ldots + c^k = c^k(1 + c + \cdots + c^{l-k-1}) < \frac{c^k}{1-c}.$$

Therefore, for all $l > k \geq 0$,

$$\|\mathbf{x}_l - \mathbf{x}_k\| < \frac{c^k}{1-c}\|\mathbf{x}_1 - \mathbf{x}_0\|.$$

Given $\varepsilon > 0$, there exists a positive integer K such that for all $k \geq K$,

$$\frac{c^k}{1-c}\|\mathbf{x}_1 - \mathbf{x}_0\| < \varepsilon.$$

This implies that for all $l > k \geq K$,

$$\|\mathbf{x}_l - \mathbf{x}_k\| < \varepsilon.$$

In other words, we have shown that $\{\mathbf{x}_k\}$ is a Cauchy sequence. Therefore, it converges to a point $\mathbf{u}$ in $\mathbb{R}^n$. Since $\mathfrak{D}$ is closed, $\mathbf{u}$ is in $\mathfrak{D}$.
Since $\mathbf{F}$ is continuous, the sequence $\{\mathbf{F}(\mathbf{x}_k)\}$ converges to $\mathbf{F}(\mathbf{u})$. But $\mathbf{F}(\mathbf{x}_k) = \mathbf{x}_{k+1}$. Being a subsequence of $\{\mathbf{x}_k\}$, the sequence $\{\mathbf{x}_{k+1}\}$ converges to $\mathbf{u}$ as well. This shows that

$$\mathbf{F}(\mathbf{u}) = \mathbf{u},$$

which says that $\mathbf{u}$ is a fixed point of $\mathbf{F}$. Now if $\mathbf{v}$ is another point in $\mathfrak{D}$ such that $\mathbf{F}(\mathbf{v}) = \mathbf{v}$, then

$$\|\mathbf{u} - \mathbf{v}\| = \|\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v})\| \leq c\|\mathbf{u} - \mathbf{v}\|.$$

Since $c \in [0, 1)$, this can only be true if $\|\mathbf{u} - \mathbf{v}\| = 0$, which implies that $\mathbf{v} = \mathbf{u}$. Hence, the fixed point of $\mathbf{F}$ is unique.

As an application of the contraction mapping theorem, we prove the following.

Theorem 2.44

Let r be a positive number and let $\mathbf{G} : B(\mathbf{0}, r) \to \mathbb{R}^n$ be a mapping such that $\mathbf{G}(\mathbf{0}) = \mathbf{0}$, and

$$\|\mathbf{G}(\mathbf{u}) - \mathbf{G}(\mathbf{v})\| \leq \frac{1}{2}\|\mathbf{u} - \mathbf{v}\| \qquad \text{for all } \mathbf{u}, \mathbf{v} \in B(\mathbf{0}, r).$$

If $\mathbf{F} : B(\mathbf{0}, r) \to \mathbb{R}^n$ is the function defined as

$$\mathbf{F}(\mathbf{x}) = \mathbf{x} + \mathbf{G}(\mathbf{x}),$$

then $\mathbf{F}$ is a one-to-one continuous mapping whose image contains the open ball $B(\mathbf{0}, r/2)$.

Proof

By definition, $\mathbf{G}$ is a contraction. Hence, it is continuous. Therefore, $\mathbf{F} : B(\mathbf{0}, r) \to \mathbb{R}^n$ is also continuous. If $\mathbf{F}(\mathbf{u}) = \mathbf{F}(\mathbf{v})$, then

$$\mathbf{u} - \mathbf{v} = \mathbf{G}(\mathbf{v}) - \mathbf{G}(\mathbf{u}).$$

Therefore,

$$\|\mathbf{u} - \mathbf{v}\| = \|\mathbf{G}(\mathbf{v}) - \mathbf{G}(\mathbf{u})\| \leq \frac{1}{2}\|\mathbf{u} - \mathbf{v}\|.$$

This implies that $\|\mathbf{u} - \mathbf{v}\| = 0$, and thus, $\mathbf{u} = \mathbf{v}$. Hence, $\mathbf{F}$ is one-to-one. Given $\mathbf{y} \in B(\mathbf{0}, r/2)$, let $r_1 = 2\|\mathbf{y}\|$. Then $r_1 < r$. Consider the map $\mathbf{H} : CB(\mathbf{0}, r_1) \to \mathbb{R}^n$ defined as

$$\mathbf{H}(\mathbf{x}) = \mathbf{y} - \mathbf{G}(\mathbf{x}).$$

For any $\mathbf{u}$ and $\mathbf{v}$ in $CB(\mathbf{0}, r_1)$,

$$\|\mathbf{H}(\mathbf{u}) - \mathbf{H}(\mathbf{v})\| = \|\mathbf{G}(\mathbf{u}) - \mathbf{G}(\mathbf{v})\| \leq \frac{1}{2}\|\mathbf{u} - \mathbf{v}\|.$$

Therefore, $\mathbf{H}$ is also a contraction. Notice that if $\mathbf{x} \in CB(\mathbf{0}, r_1)$,

$$\|\mathbf{H}(\mathbf{x})\| \leq \|\mathbf{y}\| + \|\mathbf{G}(\mathbf{x}) - \mathbf{G}(\mathbf{0})\| \leq \frac{r_1}{2} + \frac{1}{2}\|\mathbf{x}\| \leq \frac{r_1}{2} + \frac{r_1}{2} = r_1.$$

Therefore, $\mathbf{H}$ is a contraction that maps the closed set $CB(\mathbf{0}, r_1)$ into itself. By the contraction mapping theorem, there exists $\mathbf{u}$ in $CB(\mathbf{0}, r_1)$ such that $\mathbf{H}(\mathbf{u}) = \mathbf{u}$. This gives

$$\mathbf{y} - \mathbf{G}(\mathbf{u}) = \mathbf{u},$$

or equivalently,

$$\mathbf{y} = \mathbf{u} + \mathbf{G}(\mathbf{u}) = \mathbf{F}(\mathbf{u}).$$

In other words, we have shown that there exists $\mathbf{u} \in CB(\mathbf{0}, r_1) \subset B(\mathbf{0}, r)$ such that $\mathbf{F}(\mathbf{u}) = \mathbf{y}$. This proves that the image of the map $\mathbf{F} : B(\mathbf{0}, r) \to \mathbb{R}^n$ contains the open ball $B(\mathbf{0}, r/2)$.

Exercises 2.5

Question 1

Let

$$S^n = \left\{(x_1, \ldots, x_n, x_{n+1}) \in \mathbb{R}^{n+1} \,|\, x_1^2 + \cdots + x_n^2 + x_{n+1}^2 = 1\right\}$$

be the n-sphere, and let $\mathbf{F} : S^n \to S^n$ be a mapping such that

$$\|\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v})\| \leq \frac{2}{3}\|\mathbf{u} - \mathbf{v}\| \qquad \text{for all } \mathbf{u}, \mathbf{v} \in S^n.$$

Show that there is a unique $\mathbf{w} \in S^n$ such that $\mathbf{F}(\mathbf{w}) = \mathbf{w}$.

Question 2

Let r be a positive number, and let c be a positive number less than 1. Assume that $\mathbf{G} : B(\mathbf{0}, r) \to \mathbb{R}^n$ is a mapping such that $\mathbf{G}(\mathbf{0}) = \mathbf{0}$, and

$$\|\mathbf{G}(\mathbf{u}) - \mathbf{G}(\mathbf{v})\| \leq c\|\mathbf{u} - \mathbf{v}\| \qquad \text{for all } \mathbf{u}, \mathbf{v} \in B(\mathbf{0}, r).$$

If $\mathbf{F} : B(\mathbf{0}, r) \to \mathbb{R}^n$ is the function defined as

$$\mathbf{F}(\mathbf{x}) = \mathbf{x} + \mathbf{G}(\mathbf{x}),$$

show that $\mathbf{F}$ is a one-to-one continuous mapping whose image contains the open ball $B(\mathbf{0}, ar)$, where $a = 1 - c$.

Chapter 3

Continuous Functions on Connected Sets and Compact Sets

In volume I, we have seen that intermediate value theorem and extreme value theorem play important roles in analysis. In order to extend these two theorems to multivariable functions, we need to consider two topological properties of sets – the connectedness and compactness.

3.1 Path-Connectedness and Intermediate Value Theorem

We want to extend the intermediate value theorem to multivariable functions. For this, we need to consider a topological property called connectedness. In this section, we will discuss the topological property called path-connectedness first, which is a more natural concept.

Definition 3.1 Path

Let S be a subset of $\mathbb{R}^n$, and let $\mathbf{u}$ and $\mathbf{v}$ be two points in S. A *path in* S *joining* $\mathbf{u}$ *to* $\mathbf{v}$ is a **continuous** function $\boldsymbol{\gamma} : [a, b] \to S$ such that $\boldsymbol{\gamma}(a) = \mathbf{u}$ and $\boldsymbol{\gamma}(b) = \mathbf{v}$.

For any real numbers a and b with $a < b$, the map $u : [0, 1] \to [a, b]$ defined by

$$u(t) = a + t(b - a)$$

is a continuous bijection. Its inverse $u^{-1} : [a, b] \to [0, 1]$ is

$$u^{-1}(t) = \frac{t - a}{b - a},$$

which is also continuous. Hence, in the definition of a path, we can let the domain be any $[a, b]$ with $a < b$.

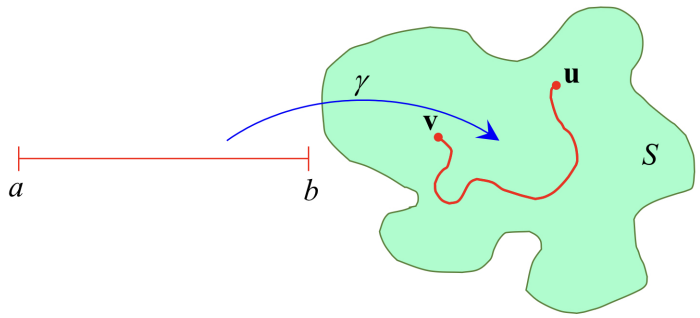


Figure 3.1: A path in S joining $\mathbf{u}$ to $\mathbf{v}$.

Example 3.1

Given a set S and a point $\mathbf{x}_0$ in S, the constant function $\boldsymbol{\gamma} : [a, b] \to S$, $\boldsymbol{\gamma}(t) = \mathbf{x}_0$, is a path in S.

If $\boldsymbol{\gamma} : [a, b] \to S$ is a path in $S \subset \mathbb{R}^n$, and S' is any other subset of $\mathbb{R}^n$ that contains the image of $\boldsymbol{\gamma}$, then $\boldsymbol{\gamma}$ is also a path in S'.

Example 3.2

Let R be the rectangle $R = [-2, 2] \times [-2, 2]$. The function $\boldsymbol{\gamma} : [0, 1] \to \mathbb{R}^2$, $\boldsymbol{\gamma}(t) = (\cos(\pi t), \sin(\pi t))$ is a path in R joining $\mathbf{u} = (1, 0)$ to $\mathbf{v} = (-1, 0)$.

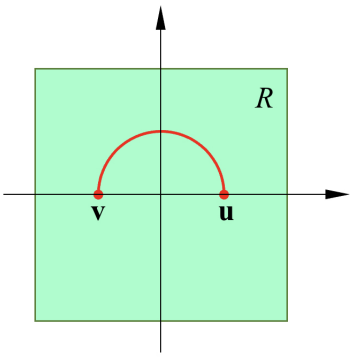


Figure 3.2: The path in Example 3.2.

Example 3.3

Let S be a subset of $\mathbb{R}^n$. If $\boldsymbol{\gamma} : [a, b] \to S$ is a path in S joining $\mathbf{u}$ to $\mathbf{v}$, then $\widetilde{\boldsymbol{\gamma}} : [-b, -a] \to S$, $\widetilde{\boldsymbol{\gamma}}(t) = \boldsymbol{\gamma}(-t)$, is a path in S joining $\mathbf{v}$ to $\mathbf{u}$.

Now we define path-connectedness.

Definition 3.2 Path-Connected

Let S be a subset of $\mathbb{R}^n$. We say that S is path-connected if any two points $\mathbf{u}$ and $\mathbf{v}$ in S can be joined by a path in S.

It is easy to characterize a path-connected subset of $\mathbb{R}$. In volume I, we have defined the concept of convex sets. A subset S of $\mathbb{R}$ is a convex set provided that for any u and v in S and any $t \in [0, 1]$, $(1-t)u+tv$ is also in S. This is equivalent to if u and v are points in S with $u < v$, all the points w satisfying $u < w < v$ is also in S. We have shown that a subset S of $\mathbb{R}$ is a convex set if and only if it is an interval.

The following theorem characterize a path-connected subset of $\mathbb{R}$.

Theorem 3.1

Let S be a subset of $\mathbb{R}$. Then S is path-connected if and only if S is an interval.

Proof

If S is an interval, then for any u and v in S, and for any $t \in [0, 1]$, $(1 - t)u + tv$ is in S. Hence, the function $\gamma : [0, 1] \to S$, $\gamma(t) = (1 - t)u + tv$ is a path in S that joins u to v.

Conversely, assume that S is a path-connected subset of $\mathbb{R}$. To show that S is an interval, we need to show that for any u and v in S with $u < v$, any w that is in the interval $[u, v]$ is also in S. Since S is path-connected, there is a path $\gamma : [0, 1] \to S$ such that $\gamma(0) = u$ and $\gamma(1) = v$. Since γ is continuous, and w is in between $\gamma(0)$ and $\gamma(1)$, intermediate value theorem implies that there is a $c \in [0, 1]$ so that $\gamma(c) = w$. Thus, w is in S.

To explore path-connected subsets of $\mathbb{R}^n$ with $n \geq 2$, we first extend the

concept of convex sets to $\mathbb{R}^n$. Given two points $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$, when t runs through all the points in the interval $[0, 1]$, $(1-t)\mathbf{u} + t\mathbf{v}$ describes all the points on the line segment between $\mathbf{u}$ and $\mathbf{v}$.

Definition 3.3 Convex Sets

Let S be a subset of $\mathbb{R}^n$. We say that S is convex if for any two points $\mathbf{u}$ and $\mathbf{v}$ in S, the line segment between $\mathbf{u}$ and $\mathbf{v}$ lies entirely in S. Equivalently, S is convex provided that for any two points $\mathbf{u}$ and $\mathbf{v}$ in S, the point $(1-t)\mathbf{u} + t\mathbf{v}$ is in S for any $t \in [0, 1]$.

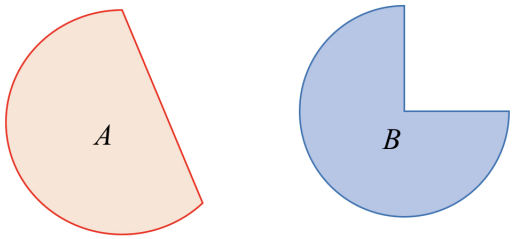


Figure 3.3: A is a convex set, B is not.

If $\mathbf{u} = (u_1, \ldots, u_n)$ and $\mathbf{v} = (v_1, \ldots, v_n)$ are two points in $\mathbb{R}^n$, the map $\boldsymbol{\gamma} : [0, 1] \to \mathbb{R}^n$,

$$\boldsymbol{\gamma}(t) = (1-t)\mathbf{u} + t\mathbf{v} = ((1-t)u_1 + tv_1, \ldots, (1-t)u_n + tv_n)$$

is a continuous functions, since each of its components is continuous. Thus, we have the following.

Theorem 3.2

Let S be a subset of $\mathbb{R}^n$. If S is convex, then it is path-connected.

Let us look at some examples of convex sets.

Example 3.4

Let $I_1, \ldots, I_n$ be intervals in $\mathbb{R}$. Show that the set $S = I_1 \times \cdots \times I_n$ is path-connected.

Solution

We claim that S is convex. Then Theorem 3.2 implies that S is path-connected.
Given that $\mathbf{u} = (u_1, \ldots, u_n)$ and $\mathbf{v} = (v_1, \ldots, v_n)$ are two points in S, for each $1 \leq i \leq n$, u_i and v_i are in I_i. Since I_i is an interval, for any $t \in [0, 1]$, $(1-t)u_i + tv_i$ is in I_i. Hence,

$$(1-t)\mathbf{u} + t\mathbf{v} = ((1-t)u_1 + tv_1, \ldots, (1-t)u_n + tv_n)$$

is in S. This shows that S is convex.

Special cases of sets of the form $S = I_1 \times \cdots \times I_n$ are open and closed rectangles.

Example 3.5

An open rectangle

$$U = (a_1, b_1) \times \cdots \times (a_n, b_n)$$

and its closure

$$R = [a_1, b_1] \times \cdots \times [a_n, b_n]$$

are convex sets. Hence, they are path-connected.

Example 3.6

Let $\mathbf{x}_0$ be a point in $\mathbb{R}^n$, and let r be a positive number. Show that the open ball $B(\mathbf{x}_0, r)$ and the closed ball $CB(\mathbf{x}_0, r)$ are path-connected sets.

Solution

Let $\mathbf{u}$ and $\mathbf{v}$ be two points in $B(\mathbf{x}_0, r)$. Then $\|\mathbf{u}-\mathbf{x}_0\| < r$ and $\|\mathbf{v}-\mathbf{x}_0\| < r$. For any $t \in [0, 1]$, $t \geq 0$ and $1 - t \geq 0$. By triangle inequality,

$$\begin{aligned} \|(1-t)\mathbf{u} + t\mathbf{v} - \mathbf{x}_0\| &\leq \|(1-t)(\mathbf{u} - \mathbf{x}_0)\| + \|t(\mathbf{v} - \mathbf{x}_0)\| \\ &= (1-t)\|\mathbf{u} - \mathbf{x}_0\| + t\|\mathbf{v} - \mathbf{x}_0\| \\ &< (1-t)r + tr = r. \end{aligned}$$

This shows that $(1-t)\mathbf{u}+t\mathbf{v}$ is in $B(\mathbf{x}_0, r)$. Hence, $B(\mathbf{x}_0, r)$ is convex. Replacing $<$ by $\leq$, one can show that $CB(\mathbf{x}_0, r)$ is convex.
By Theorem 3.2, the open ball $B(\mathbf{x}_0, r)$ and the closed ball $CB(\mathbf{x}_0, r)$ are path-connected sets.

Not all the path-connected sets are convex. Before we give an example, let us first prove the following useful lemma.

Lemma 3.3

Let A and B be path-connected subsets of $\mathbb{R}^n$. If $A \cap B$ is nonempty, then $S = A \cup B$ is path-connected.

Proof

Let $\mathbf{u}$ and $\mathbf{v}$ be two points in S. If both $\mathbf{u}$ and $\mathbf{v}$ are in the set A, then they can be joined by a path in A, which is also in S. Similarly, if both $\mathbf{u}$ and $\mathbf{v}$ are in the set B, then they can be joined by a path in S. If $\mathbf{u}$ is in A and $\mathbf{v}$ is in B, let $\mathbf{x}_0$ be any point in $A \cap B$. Then $\mathbf{u}$ and $\mathbf{x}_0$ are both in the path-connected set A, and $\mathbf{v}$ and $\mathbf{x}_0$ are both in the path-connected set B. Therefore, there exist continuous functions $\boldsymbol{\gamma}_1 : [0,1] \to A$ and $\boldsymbol{\gamma}_2 : [1,2] \to B$ such that $\boldsymbol{\gamma}_1(0) = \mathbf{u}$, $\boldsymbol{\gamma}_1(1) = \mathbf{x}_0$, $\boldsymbol{\gamma}_2(1) = \mathbf{x}_0$ and $\boldsymbol{\gamma}_2(2) = \mathbf{v}$. Define the function $\boldsymbol{\gamma} : [0,2] \to A \cup B$ by

$$\boldsymbol{\gamma}(t) = \begin{cases} \boldsymbol{\gamma}_1(t), & \text{if } 0 \leq t \leq 1, \\ \boldsymbol{\gamma}_2(t), & \text{if } 1 \leq t \leq 2. \end{cases}$$

Since $[0,1]$ and $[1,2]$ are closed subsets of $\mathbb{R}$, the function $\boldsymbol{\gamma} : [0,2] \to S$ is continuous. Thus, $\boldsymbol{\gamma}$ is a path in S from $\mathbf{u}$ to $\mathbf{v}$. This proves that S is path-connected.

Now we can give an example of a path-connected set that is not convex.

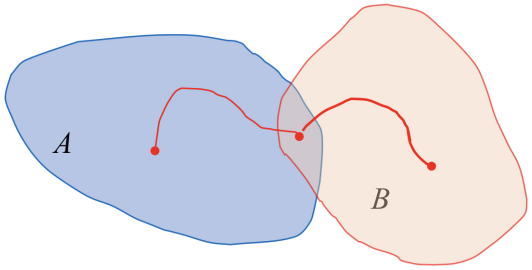


Figure 3.4: If two sets A and B are path-connected and $A \cap B$ is nonempty, then $A \cup B$ is also path-connected.

Example 3.7

Show that the set

$$S = \{(x, y) \,|\, 0 \leq x \leq 1, -2 \leq y \leq 2\} \cup \left\{(x, y) \,|\, (x-2)^2 + y^2 \leq 1\right\}$$

is path-connected, but not convex.

Solution

The set

$$A = \{(x, y) \,|\, 0 \leq x \leq 1, -2 \leq y \leq 2\} = [0, 1] \times [-2, 2]$$

is a closed rectangle. Therefore, it is path- connected. The set

$$B = \left\{(x, y) \,|\, (x-2)^2 + y^2 \leq 1\right\}$$

is a closed ball with center at $(2, 0)$ and radius 1. Hence, it is also path-connected. Since the point $\mathbf{x}_0 = (1, 0)$ is in both A and B, $S = A \cup B$ is path-connected.
The points $\mathbf{u} = (1, 2)$ and $\mathbf{v} = (2, 1)$ are in S. Consider the point

$$\mathbf{w} = \frac{1}{2}\mathbf{u} + \frac{1}{2}\mathbf{v} = \left(\frac{3}{2}, \frac{3}{2}\right).$$

It is not in S. This shows that S is not convex.

Let us now prove the following important theorem which says that continuous

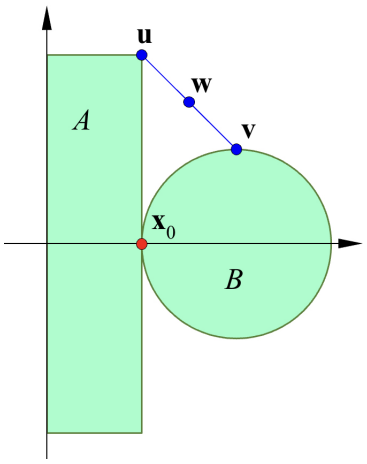


Figure 3.5: The set $A \cup B$ is path-connected but not convex.

functions preserve path-connectedness.

Theorem 3.4

Let $\mathfrak{D}$ be a path-connected subset of $\mathbb{R}^n$. If $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is a continuous function, then $\mathbf{F}(\mathfrak{D})$ is path-connected.

Proof

Let $\mathbf{v}_1$ and $\mathbf{v}_2$ be two points in $\mathbf{F}(\mathfrak{D})$. Then there exist $\mathbf{u}_1$ and $\mathbf{u}_2$ in $\mathfrak{D}$ such that $\mathbf{F}(\mathbf{u}_1) = \mathbf{v}_1$ and $\mathbf{F}(\mathbf{u}_2) = \mathbf{v}_2$. Since $\mathfrak{D}$ is path-connected, there is a continuous function $\boldsymbol{\gamma} : [0, 1] \to \mathfrak{D}$ such that $\boldsymbol{\gamma}(0) = \mathbf{u}_1$ and $\boldsymbol{\gamma}(1) = \mathbf{u}_2$. The map $\boldsymbol{\alpha} = (\mathbf{F} \circ \boldsymbol{\gamma}) : [0, 1] \to \mathbf{F}(\mathfrak{D})$ is then a conitnuous map with $\boldsymbol{\alpha}(0) = \mathbf{v}_1$ and $\boldsymbol{\alpha}(1) = \mathbf{v}_2$. This shows that $\mathbf{F}(\mathfrak{D})$ is path-connected.

From Theorem 3.4, we obtain the following.

Theorem 3.5 Intermediate Value Theorem for Path-Connected Sets

Let $\mathfrak{D}$ be a path-connected subset of $\mathbb{R}^n$, and let $f : \mathfrak{D} \to \mathbb{R}$ be a function defined on $\mathfrak{D}$. If f is continuous, then $f(\mathfrak{D})$ is an interval.

Proof

By Theorem 3.4, $f(\mathfrak{D})$ is a path-connected subset of $\mathbb{R}$. By Theorem 3.1, $f(\mathfrak{D})$ is an interval.

We can also use Theorem 3.4 to establish more examples of path-connected sets. Let us first look at an example.

Example 3.8

Show that the circle

$$S^1 = \{(x, y) \mid x^2 + y^2 = 1\}$$

is path-connected.

Solution

Define the function $f : [0, 2\pi] \to \mathbb{R}^2$ by

$$f(t) = (\cos t, \sin t).$$

Notice that $S^1 = f([0, 2\pi))$. Since each component of f is a continuous function, f is a continuous function. Since $[0, 2\pi]$ is an interval, it is path-connected. By Theorem 3.4, $S^1 = f([0, 2\pi])$ is path-connected.

A more general theorem is as follows.

Theorem 3.6

Let $\mathfrak{D}$ be a path-connected subset of $\mathbb{R}^n$, and let $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ be a function defined on $\mathfrak{D}$. If $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is continuous, then the graph of $\mathbf{F}$,

$$G_{\mathbf{F}} = \{(\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \in \mathfrak{D}, \mathbf{y} = \mathbf{F}(\mathbf{x})\}$$

is a path-connected subset of $\mathbb{R}^{n+m}$.

Proof

By Corollary 2.32, the function $\mathbf{H} : \mathfrak{D} \to \mathbb{R}^{n+m}$, $\mathbf{H}(\mathbf{x}) = (\mathbf{x}, \mathbf{F}(\mathbf{x}))$, is continuous. Since $\mathbf{H}(\mathfrak{D}) = G_{\mathbf{F}}$, Theorem 3.4 implies that $G_{\mathbf{F}}$ is a path-connected subset of $\mathbb{R}^{n+m}$.

Now let us consider spheres, which are boundary of balls.

Definition 3.4 The Standard Unit n-Sphere S^n

A standard unit n-sphere S^n is a subset of $\mathbb{R}^{n+1}$ consists of all points $\mathbf{x} = (x_1, \ldots, x_n, x_{n+1})$ in $\mathbb{R}^{n+1}$ satisfying the equation $\|\mathbf{x}\| = 1$, namely,

$$x_1^2 + \cdots + x_n^2 + x_{n+1}^2 = 1.$$

The n-sphere S^n is the boundary of the $(n+1)$ open ball $B^{n+1} = B(\mathbf{0}, 1)$ with center at the origin and radius 1.

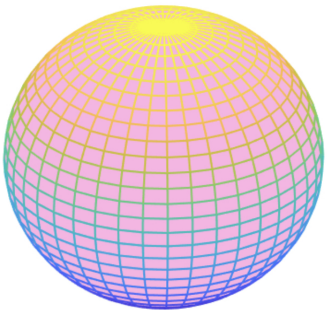


Figure 3.6: A sphere.

Example 3.9

Show that the standard unit n-sphere S^n is path-connected.

Solution

Notice that $S^n = S^n_+ \cup S^n_-$, where S^n_+ and S^n_- are respectively the upper and lower hemispheres with $x_{n+1} \geq 0$ and $x_{n+1} \leq 0$ respectively.

If $\mathbf{x} \in S^n_+$, then

$$x_{n+1} = \sqrt{1 - x_1^2 - \ldots - x_n^2};$$

whereas if $\mathbf{x} \in S^n_-$,

$$x_{n+1} = -\sqrt{1 - x_1^2 - \ldots - x_n^2}.$$

Let

$$CB^n = \left\{(x_1, \ldots, x_n) \,|\, x_1^2 + \cdots + x_n^2 \leq 1\right\}$$

be the closed ball in $\mathbb{R}^n$ with center at the origin and radius 1. Define the functions $f_\pm : CB^n \to \mathbb{R}$ by

$$f_\pm(x_1, \ldots, x_n) = \pm\sqrt{1 - x_1^2 - \ldots - x_n^2}.$$

Notice that S^n_+ and S^n_- are respectively the graphs of f_+ and f_-. Since they are compositions of the square root function and a polynomial function, which are both continuous, f_+ and f_- are continuous functions. The closed ball CB^n is path-connected. Theorem 3.6 then implies that S^n_+ and S^n_- are path-connected.

Since both S^n_+ and S^n_- contain the unit vector $\mathbf{e}_1$ in $\mathbb{R}^{n+1}$, the set $S^n_+ \cap S^n_-$ is nonempty. By Lemma 3.3, $S^n = S^n_+ \cup S^n_-$ is path-connected.

Remark 3.1

There is an alternative way to prove that the n-sphere S^n is path-connected. Given two distinct points $\mathbf{u}$ and $\mathbf{v}$ in S^n, they are unit vectors in $\mathbb{R}^{n+1}$. We want to show that there is a path in S^n joining $\mathbf{u}$ to $\mathbf{v}$.

Notice that the line segment $L = \{(1-t)\mathbf{u} + t\mathbf{v} \,|\, 0 \leq t \leq 1\}$ in $\mathbb{R}^{n+1}$ contains the origin if and only if $\mathbf{u}$ and $\mathbf{v}$ are parallel, if and only if $\mathbf{v} = -\mathbf{u}$. Thus, we discuss two cases.

Case I: $\mathbf{v} \neq -\mathbf{u}$.
In this case, let $\boldsymbol{\gamma} : [0, 1] \to \mathbb{R}^{n+1}$ be the function defined as

$$\boldsymbol{\gamma}(t) = \frac{(1-t)\mathbf{u} + t\mathbf{v}}{\|(1-t)\mathbf{u} + t\mathbf{v}\|}.$$

Since $(1-t)\mathbf{u} + t\mathbf{v} \neq \mathbf{0}$ for all $0 \leq t \leq 1$, $\boldsymbol{\gamma}$ is a continuous function. It is easy to check that its image lies in S^n. Hence, $\boldsymbol{\gamma}$ is a path in S^n joining $\mathbf{u}$ to $\mathbf{v}$.
Case 2: $\mathbf{v} = -\mathbf{u}$.
In this case, let $\mathbf{w}$ be a unit vector orthogonal to $\mathbf{u}$, and let $\boldsymbol{\gamma} : [0, \pi] \to \mathbb{R}^{n+1}$ be the function defined as

$$\boldsymbol{\gamma}(t) = (\cos t)\mathbf{u} + (\sin t)\mathbf{w}.$$

Since $\sin t$ and $\cos t$ are continuous functions, $\boldsymbol{\gamma}$ is a continuous function. Since $\mathbf{u}$ and $\mathbf{w}$ are orthogonal, the generalized Pythagoras theorem implies that

$$\|\boldsymbol{\gamma}(t)\|^2 = \cos^2 t\|\mathbf{u}\|^2 + \sin^2 t\|\mathbf{w}\|^2 = \cos^2 t + \sin^2 t = 1.$$

Therefore, the image of $\boldsymbol{\gamma}$ lies in S^n. It is easy to see that $\boldsymbol{\gamma}(0) = \mathbf{u}$ and $\boldsymbol{\gamma}(\pi) = -\mathbf{u} = \mathbf{v}$. Hence, $\boldsymbol{\gamma}$ is a path in S^n joining $\mathbf{u}$ to $\mathbf{v}$.

Example 3.10

Let $f : S^n \to \mathbb{R}$ be a continuous function. Show that there is a point $\mathbf{u}_0$ on S^n such that $f(\mathbf{u}_0) = f(-\mathbf{u}_0)$.

Solution

The function $\mathbf{g} : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$, $\mathbf{g}(\mathbf{u}) = -\mathbf{u}$ is a linear transformation. Hence, it is continuous. Restricted to S^n, $\mathbf{g}(S^n) = S^n$. Thus, the function $f_1 : S^n \to \mathbb{R}$, $f_1(\mathbf{u}) = f(-\mathbf{u})$, is also continuous.

It follows that the function $h : S^n \to \mathbb{R}$ defined by

$$h(\mathbf{u}) = f(\mathbf{u}) - f(-\mathbf{u})$$

is continuous. Notice that

$$h(-\mathbf{u}) = f(-\mathbf{u}) - f(\mathbf{u}) = -h(\mathbf{u}).$$

This implies that if the number a is in the range of h, so does the number $-a$. Since the number 0 is in between a and $-a$ for any a, and S^n is path-connected, intermediate value theorem implies that the number 0 is also in the range of h. This means that there is an $\mathbf{u}_0$ on S^n such that $h(\mathbf{u}_0) = 0$. Equivalently, $f(\mathbf{u}_0) = f(-\mathbf{u}_0)$.

Theorem 3.5 says that a continuous function defined on a path-connected set satisfies the intermediate value theorem. We make the following definition.

Definition 3.5 Intermediate Value Property

Let S be a subset of $\mathbb{R}^n$. We say that S has intermediate value property provided that whenever $f : S \to \mathbb{R}$ is a continuous function, then $f(S)$ is an interval.

Theorem 3.5 says that if S is a path-connected set, then it has intermediate value property. It is natural to ask whether it is true that any set S that has the intermediate value property must be path-connected. Unfortunately, it turns out that the answer is yes only when S is a subset of $\mathbb{R}$. If S is a subset of $\mathbb{R}^n$ with $n \geq 2$, this is not true. This leads us to define a new property of sets called connectedness in the next section.

Exercises 3.1

Question 1

Is the set $A = (-1, 2) \cup (2, 5]$ path-connected? Justify your answer.

Question 2

Let a and b be positive numbers, and let A be the subset of $\mathbb{R}^2$ given by

$$A = \left\{ (x, y) \;\middle|\; \frac{x^2}{a^2} + \frac{y^2}{b^2} \leq 1 \right\}.$$

Show that A is convex, and deduce that it is path-connected.

Question 3

Let (a, b, c) be a nonzero vector, and let $\mathbb{P}$ be the plane in $\mathbb{R}^3$ given by

$$\mathbb{P} = \{(x, y, z) \,|\, ax + by + cz = d\},$$

where d is a constant. Show that $\mathbb{P}$ is convex, and deduce that it is path-connected.

Question 4

Let S be the subset of $\mathbb{R}^3$ given by

$$S = \{(x, y, z) \,|\, x > 0, y \leq 1, 2 \leq z < 7\}.$$

Show that S is path-connected.

Question 5

Let a, b and c be positive numbers, and let S be the subset of $\mathbb{R}^3$ given by

$$S = \left\{ (x, y, z) \;\middle|\; \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 \right\}.$$

Show that S is path-connected.

Question 6

Let $\mathbf{u} = (3, 0)$ and let A be the subset of $\mathbb{R}^2$ given by

$$A = \left\{(x, y) \,|\, x^2 + y^2 \leq 1\right\}.$$

Define the function $f : A \to \mathbb{R}$ by $f(\mathbf{x}) = d(\mathbf{x}, \mathbf{u})$.

(a) Find $f(\mathbf{x}_1)$ and $f(\mathbf{x}_2)$, where $\mathbf{x}_1 = (1, 0)$ and $\mathbf{x}_2 = (-1, 0)$.

(b) Use intermediate value theorem to justify that there is a point $\mathbf{x}_0$ in A such that $d(\mathbf{x}_0, \mathbf{u}) = \pi$.

Question 7

Let A and B be subsets of $\mathbb{R}^n$. If A and B are convex, show that $A \cap B$ is also convex.

3.2 Connectedness and Intermediate Value Property

In this section, we study a property of sets which is known as connectedness. Let us first look at the path-connected subsets of $\mathbb{R}$ from a different perpective. We have shown in the previous section that a subset of $\mathbb{R}$ is path-connected if and only if it is an interval. A set of the form

$$A = (-2, 2] \setminus \{0\} = (-2, 0) \cup (0, 2]$$

is not path-connected, since it contains the points -1 and 1, but it does not contain the point 0 that is in between. Intuitively, there is no way to go from the point -1 to 1 *continuously* without leaving the set A.

Let $U = (-\infty, 0)$ and $V = (0, \infty)$. Notice that U and V are open subsets of $\mathbb{R}$ which both intersect the set A. Moreover,

$$A = (A \cap U) \cup (A \cap V),$$

or equivalently,

$$A \subset U \cup V.$$

We say that A is *separated* by the open sets U and V.

Definition 3.6 Separation of a Set

Let A be a subset of $\mathbb{R}^n$. A *separation* of A is a pair (U, V) of subsets of $\mathbb{R}^n$ which satisfies the following conditions.

(a) U and V are open sets.

(b) $A \cap U \neq \emptyset$ and $A \cap V \neq \emptyset$.

(c) $A \subset U \cup V$, or equivalently, A is the union of $A \cap U$ and $A \cap V$.

(d) A is disjoint from $U \cap V$, or equivalently, $A \cap U$ and $A \cap V$ are disjoint.

If (U, V) is a separation of A, we say that A is separated by the open sets U and V, or the open sets U and V separate A.

Example 3.11

Let $A = (-2, 0) \cup (0, 2]$, and let $U = (-\infty, 0)$ and $V = (0, \infty)$. Then the open sets U and V separate A.
Let $U_1 = (-3, 0)$ and $V_1 = (0, 3)$. The open sets U_1 and V_1 also separate A.

Now we define connectedness.

Definition 3.7 Connected Sets

Let A be a subset of $\mathbb{R}^n$. We say that A is connected if there does not exist a pair of open sets U and V that separate A.

Example 3.12

Determine whether the set

$$A = \{(x, y) \,|\, y = 0\} \cup \left\{ (x, y) \,\middle|\, y = \frac{2}{1 + x^2} \right\}$$

is connected.

Solution

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function defined as

$$f(x, y) = y(x^2 + 1).$$

Since f is a polynomial function, it is continuous. The intervals $V_1 = (-1, 1)$ and $V_2 = (1, 3)$ are open sets in $\mathbb{R}$. Hence, the sets $U_1 = f^{-1}(V_1)$ and $U_2 = f^{-1}(V_2)$ are disjoint and they are open in $\mathbb{R}^2$. Notice that

$$A \cap U_1 = \{(x, y) \,|\, y = 0\}, \qquad A \cap U_2 = \left\{ (x, y) \,\middle|\, y = \frac{2}{1 + x^2} \right\}.$$

Thus, $A \cap U_1$ and $A \cap U_2$ are nonempty, $A \cap U_1$ and $A \cap U_2$ are disjoint, and A is a union of $A \cap U_1$ and $A \cap U_2$. This shows that the open sets U_1 and U_2 separate A. Hence, A is not connected.

Now let us explore the relation between path-connected and connected. We

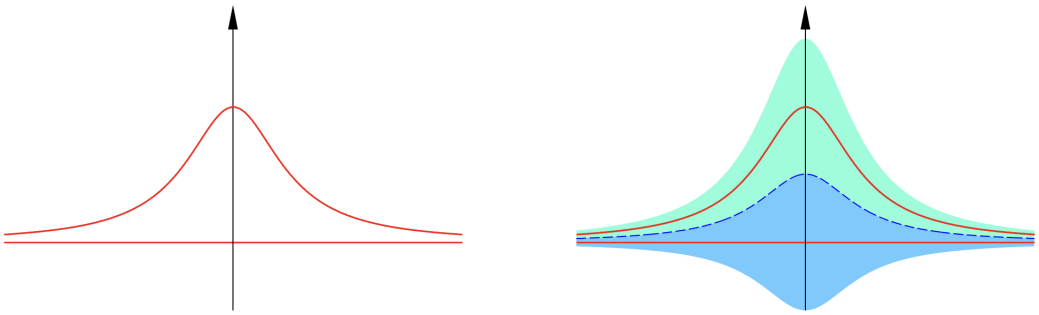


Figure 3.7: The set A defined in Example 3.12 is not connected.

first prove the following.

Theorem 3.7

Let A be a subset of $\mathbb{R}^n$, and assume that the open sets U and V separate A. Define the function $f : A \to \mathbb{R}$ by

$$f(\mathbf{x}) = \begin{cases} 0, & \text{if } \mathbf{x} \in A \cap U, \\ 1, & \text{if } \mathbf{x} \in A \cap V. \end{cases}$$

Then f is continuous.

Notice that the function f is well defined since $A \cap U$ and $A \cap V$ are disjoint.

Proof

Let $\mathbf{x}_0$ be a point in A. We want to prove that f is continuous at $\mathbf{x}_0$. Since A is contained in $U \cup V$, $\mathbf{x}_0$ is in U or in V. It suffices to consider the case where $\mathbf{x}_0$ is in U. The case where $\mathbf{x}_0$ is in V is similar.
If $\mathbf{x}_0$ is in U, since U is open, there is an $r > 0$ such that $B(\mathbf{x}_0, r) \subset U$. If $\{\mathbf{x}_k\}$ is a sequence in A that converges $\mathbf{x}_0$, there exists a positive integer K such that for all $k \geq K$, $\|\mathbf{x}_k - \mathbf{x}_0\| < r$. Thus, for all $k \geq K$, $\mathbf{x}_k \in B(\mathbf{x}_0, r) \subset U$, and hence, $f(\mathbf{x}_k) = 0$. This proves that the sequence $\{f(\mathbf{x}_k)\}$ converges to 0, which is $f(\mathbf{x}_0)$. Therefore, f is continuous at $\mathbf{x}_0$.

Now we can prove the theorem which says that a path-connected set is connected.

Theorem 3.8

Let A be a subset of $\mathbb{R}^n$. If A is path-connected, then it is connected.

Proof

We prove the contrapositive, which says that if A is not connected, then it is not path-connected.

If A is not connected, there is a pair of open sets U and V that separate A. By Theorem 3.7, the function $f : A \to \mathbb{R}$ defined by

$$f(\mathbf{x}) = \begin{cases} 0, & \text{if } \mathbf{x} \in A \cap U, \\ 1, & \text{if } \mathbf{x} \in A \cap V \end{cases}$$

is continuous. Since $f(A) = \{0, 1\}$ is not an interval, by the contrapositive of the intermediate value theorem for path-connected sets, A is not path-connected.

Theorem 3.8 provides us a large library of connected sets.

Example 3.13

The following sets are path-connected. Hence, they are also connected.

1. A set S in $\mathbb{R}^n$ of the form $S = I_1 \times \cdots \times I_n$, where $I_1, \ldots, I_n$ are intervals in $\mathbb{R}$.
2. Open rectangles and closed rectangles.
3. Open balls and closed balls.
4. The n-sphere S^n.

The following theorem says that path-connectedness and connectedness are equivalent in $\mathbb{R}$.

Theorem 3.9

Let S be a subset of $\mathbb{R}$. Then the following are equivalent.

(a) S is an interval.

(b) S is path-connected.

(c) S is connected.

Proof

We have proved (a) $\Longleftrightarrow$ (b) in the previous section. In particular, (a) implies (b). Theorem 3.8 says that (b) implies (c). Now we only need to prove that (c) implies (a).
Assume that (a) is not true. Namely, S is not an interval. Then there are points u and v in S with $u < v$, such that there is a $w \in (u, v)$ that is not in S. Let $U = (-\infty, w)$ and $V = (w, \infty)$. Then U and V are disjoint open subsets of $\mathbb{R}$. Since $w \notin S$, $S \subset U \cup V$. Since $u \in S \cap U$ and $v \in S \cap V$, $S \cap U$ and $S \cap V$ are nonempty. Hence, U and V are open sets that separate S. This shows that S is not connected. Thus, we have proved that if (a) is not true, then (c) is not true. This is equivalent to (c) implies (a).

Connectedness is also preserved by continuous functions.

Theorem 3.10

Let $\mathfrak{D}$ be a connected subset of $\mathbb{R}^n$. If $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is a continuous function, then $\mathbf{F}(\mathfrak{D})$ is connected.

Proof

We prove the contra-positive. Assume that $\mathbf{F}(\mathfrak{D})$ is not connected. Then there are open sets V_1 and V_2 in $\mathbb{R}^m$ that separate $\mathbf{F}(\mathfrak{D})$. Let

$$\mathfrak{D}_1 = \{\mathbf{x} \in \mathfrak{D} \,|\, \mathbf{F}(\mathbf{x}) \in V_1\},$$
$$\mathfrak{D}_2 = \{\mathbf{x} \in \mathfrak{D} \,|\, \mathbf{F}(\mathbf{x}) \in V_2\}.$$

Since $\mathbf{F}(\mathfrak{D}) \cap V_1$ and $\mathbf{F}(\mathfrak{D}) \cap V_2$ are nonempty, $\mathfrak{D}_1$ and $\mathfrak{D}_2$ are nonempty. Since $\mathbf{F}(\mathfrak{D}) \subset V_1 \cup V_2$, $\mathfrak{D} = \mathfrak{D}_1 \cup \mathfrak{D}_2$. Since $V_1 \cap V_2$ is disjoint from $\mathbf{F}(\mathfrak{D})$, $\mathfrak{D}_1$ and $\mathfrak{D}_2$ are disjoint. However, $\mathfrak{D}_1$ and $\mathfrak{D}_2$ are not necessary open sets. We will define two open sets U_1 and U_2 in $\mathbb{R}^n$ such that $\mathfrak{D}_1 = \mathfrak{D} \cap U_1$ and $\mathfrak{D}_2 = \mathfrak{D} \cap U_2$. Then U_1 and U_2 are open sets that separate $\mathfrak{D}$.
For each $\mathbf{x}_0$ in $\mathfrak{D}_1$, $\mathbf{F}(\mathbf{x}_0) \in V_1$. Since V_1 is open, there exists $\varepsilon_{\mathbf{x}_0} > 0$ such that the ball $B(\mathbf{F}(\mathbf{x}_0), \varepsilon_{\mathbf{x}_0})$ is contained in V_1. By the continuity of $\mathbf{F}$ at $\mathbf{x}_0$, there exists $\delta_{\mathbf{x}_0} > 0$ such that for all $\mathbf{x}$ in $\mathfrak{D}$, if $\mathbf{x} \in B(\mathbf{x}_0, \delta_{\mathbf{x}_0})$, then $\mathbf{F}(\mathbf{x}) \in B(\mathbf{F}(\mathbf{x}_0), \varepsilon_{\mathbf{x}_0}) \subset V_1$. In other words,

$$\mathfrak{D} \cap B(\mathbf{x}_0, \delta_{\mathbf{x}_0}) \subset \mathbf{F}^{-1}(V_1) = \mathfrak{D}_1.$$

Notice that $B(\mathbf{x}_0, \delta_{\mathbf{x}_0})$ is an open set. Define

$$U_1 = \bigcup_{\mathbf{x}_0 \in \mathfrak{D}_1} B(\mathbf{x}_0, \delta_{\mathbf{x}_0}).$$

Being a union of open sets, U_1 is open. Since

$$\mathfrak{D} \cap U_1 = \bigcup_{\mathbf{x}_0 \in \mathfrak{D}_1} (\mathfrak{D} \cap B(\mathbf{x}_0, \delta_{\mathbf{x}_0})) \subset \mathfrak{D}_1,$$

and

$$\mathfrak{D}_1 = \bigcup_{\mathbf{x}_0 \in \mathfrak{D}_1} \{\mathbf{x}_0\} \subset \bigcup_{\mathbf{x}_0 \in \mathfrak{D}_1} (\mathfrak{D} \cap B(\mathbf{x}_0, \delta_{\mathbf{x}_0})) = \mathfrak{D} \cap U_1,$$

we find that $\mathfrak{D} \cap U_1 = \mathfrak{D}_1$. Similarly, define

$$U_2 = \bigcup_{\mathbf{x}_0 \in \mathfrak{D}_2} B(\mathbf{x}_0, \delta_{\mathbf{x}_0}).$$

Then U_2 is an open set and $\mathfrak{D} \cap U_2 = \mathfrak{D}_2$. This completes the construction of the open sets U_1 and U_2 that separate $\mathfrak{D}$. Thus, $\mathfrak{D}$ is not connected.

From Theorem 3.9 and Theorem 3.10, we also have an intermediate value theorem for connected sets.

Theorem 3.11 Intermediate Value Theorem for Connected Sets

Let $\mathfrak{D}$ be a connected subset of $\mathbb{R}^n$, and let $f : \mathfrak{D} \to \mathbb{R}$ be a function defined on $\mathfrak{D}$. If f is continuous, then $f(\mathfrak{D})$ is an interval.

Proof

By Theorem 3.10, $f(\mathfrak{D})$ is a connected subset of $\mathbb{R}$. By Theorem 3.9, $f(\mathfrak{D})$ is an interval.

Now we can prove the following.

Theorem 3.12

Let S be a subset of $\mathbb{R}^n$. Then S is connected if and only if it has the intermediate value property.

Proof

If S is connected and $f : S \to \mathbb{R}$ is continuous, Theorem 3.11 implies that $f(S)$ is an interval. Hence, S has the intermediate value property.
If S is not connected, Theorem 3.7 gives a continuous function $f : S \to \mathbb{R}$ such that $f(S) = \{0, 1\}$ is not an interval. Thus, S does not have the intermediate value property.

To give an example of a connected set that is not path-connected, we need a lemma.

Lemma 3.13

Let A be a subset of $\mathbb{R}^n$ that is separated by the open sets U and V. If C is a connected subset of A, then $C \cap U = \emptyset$ or $C \cap V = \emptyset$.

Proof

Since $C \subset A$, $C \subset U \cup V$, and C is disjoint from $U \cap V$. If $C \cap U \neq \emptyset$ and $C \cap V \neq \emptyset$, then the open sets U and V also separate C. This contradicts to C is connected. Thus, we must have $C \cap U = \emptyset$ or $C \cap V = \emptyset$.

Theorem 3.14

Let A be a connected subset of $\mathbb{R}^n$. If B is a subset of $\mathbb{R}^n$ such that

$$A \subset B \subset \overline{A},$$

then B is also connected.

Proof

If B is not connected, there exist open sets U and V in $\mathbb{R}^n$ that separate A. Since A is connected, Lemma 3.13 says that $A \cap U = \emptyset$ or $A \cap V = \emptyset$. Without loss of generality, assume that $A \cap V = \emptyset$. Then $A \subset \mathbb{R}^n \setminus V$. Thus, $\mathbb{R}^n \setminus V$ is a closed set that contains A. This implies that $\overline{A} \subset \mathbb{R}^n \setminus V$. Hence, we also have $B \subset \mathbb{R}^n \setminus V$, which contradicts to the fact that the set $B \cap V$ is not empty.

Example 3.14 The Topologist's Sine Curve

Let S be the subset of $\mathbb{R}^2$ given by $S = A \cup L$, where

$$A = \left\{ (x, y) \,\middle|\, 0 < x \leq 1, y = \sin\left(\frac{1}{x}\right) \right\},$$

and

$$L = \{(x, y) \,|\, x = 0, -1 \leq y \leq 1\}.$$

(a) Show that $S \subset \overline{A}$.

(b) Show that S is connected.

(c) Show that S is not path-connected.

Solution

(a) Since $A \subset \overline{A}$, it suffices to show that $L \subset \overline{A}$. Given $(0, u) \in L$, $-1 \leq u \leq 1$. Thus, $a = \sin^{-1} u \in [-\pi/2, \pi/2]$. Let

$$x_k = \frac{1}{a + 2\pi k} \qquad \text{for } k \in \mathbb{Z}^+.$$

Notice that $x_k \in (0, 1]$ and

$$\sin \frac{1}{x_k} = \sin a = u.$$

Thus, $\{(x_k, \sin(1/x_k))\}$ is a sequence of points in A that converges to $(0, u)$. This proves that $(0, u) \in \overline{A}$. Hence, $L \subset \overline{A}$.

(b) The interval $(0, 1]$ is path-connected and the function $f : (0, 1] \to \mathbb{R}$, $f(x) = \sin\left(\frac{1}{x}\right)$ is continuous. Thus, $A = G_f$ is path-connected, and hence it is connected. Since $A \subset S \subset \overline{A}$, Theorem 3.14 implies that S is connected.

(c) If S is path connected, there is a path $\boldsymbol{\gamma} : [0, 1] \to S$ such that $\boldsymbol{\gamma}(0) = (0, 0)$ and $\boldsymbol{\gamma}(1) = (1, \sin 1)$. Let $\boldsymbol{\gamma}(t) = (\gamma_1(t), \gamma_2(t))$. Then $\gamma_1 : [0, 1] \to \mathbb{R}$ and $\gamma_2 : [0, 1] \to \mathbb{R}$ are continuous functions. Consider the sequence $\{x_k\}$ with

$$x_k = \frac{1}{\frac{\pi}{2} + \pi k}, \qquad k \in \mathbb{Z}^+.$$

Notice that $\{x_k\}$ is a decreasing sequence of points in $[0, 1]$ that converges to 0. For each $k \in \mathbb{Z}^+$, $(x_k, y_k) \in S$ if and onlly if $y_k = \sin(1/x_k)$.

Since $\gamma_1 : [0, 1] \to \mathbb{R}$ is continuous, $\gamma_1(0) = 0$ and $\gamma_1(1) = 1$, intermediate value theorem implies that there exists $t_1 \in [0, 1]$ such that $\gamma_1(t_1) = x_1$. Similarly, there exists $t_2 \in [0, t_1]$ such that $\gamma_1(t_2) = x_2$. Continue the argument gives a decreasing sequence $\{t_k\}$ in $[0, 1]$ such that $\gamma_1(t_k) = x_k$ for all $k \in \mathbb{Z}^+$. Since the sequence $\{t_k\}$ is bounded below, it converges to some t_0 in $[0, 1]$. Since $\gamma_2 : [0, 1] \to \mathbb{R}$ is also continuous, the sequence $\{\gamma_2(t_k)\}$ should converge to $\gamma_2(t_0)$.

Since $\boldsymbol{\gamma}(t_k) \in S$ and $\gamma_1(t_k) = x_k$, we must have $\gamma_2(t_k) = y_k = (-1)^k$. But then the sequence $\{\gamma_2(t_k)\}$ is not convergent. This gives a contradiction. Hence, there does not exist a path in S that joins the point $(0, 0)$ to the point $(1, \sin 1)$. This proves that S is not path-connected.

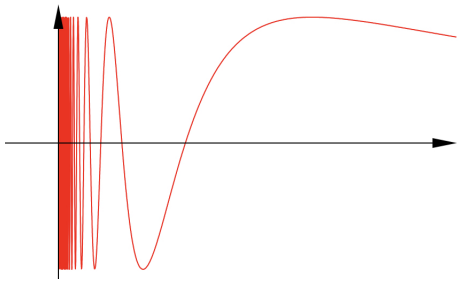


Figure 3.8: The topologist's sine curve.

Remark 3.2

Example 3.14 gives a set that is connected but not path-connected.

1. One can in fact show that $S = \overline{A}$.
2. To show that A is connected, we can also use the fact that if $\mathfrak{D}$ is a connected subset of $\mathbb{R}^n$, and $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is a continuous function, then the graph of $\mathbf{F}$ is connected. The proof of this fact is left as an exercise.

At the end of this section, we want to give a sufficient condition for a connected subset of $\mathbb{R}^n$ to be path-connected.

First we define the meaning of a polygonal path.

Definition 3.8 Polygonal Path

Let S be a subset of $\mathbb{R}^n$, and let $\mathbf{u}$ and $\mathbf{v}$ be two points in S. A path $\boldsymbol{\gamma} : [a, b] \to S$ in S that joins $\mathbf{u}$ to $\mathbf{v}$ is a polygonal path provided that there is a partition $P = \{t_0, t_1, \ldots, t_k\}$ of $[a, b]$ such that for $1 \leq i \leq k$,

$$\boldsymbol{\gamma}(t) = \mathbf{x}_{i-1} + \frac{t - t_{i-1}}{t_i - t_{i-1}} (\mathbf{x}_i - \mathbf{x}_{i-1}), \quad \text{when } t_{i-1} \leq t \leq t_i.$$

Obviously, we have the following.

Proposition 3.15

If S is a convex subset of $\mathbb{R}^n$, then any two points in S can be joined by a polygonal path in $\mathbb{R}^n$.

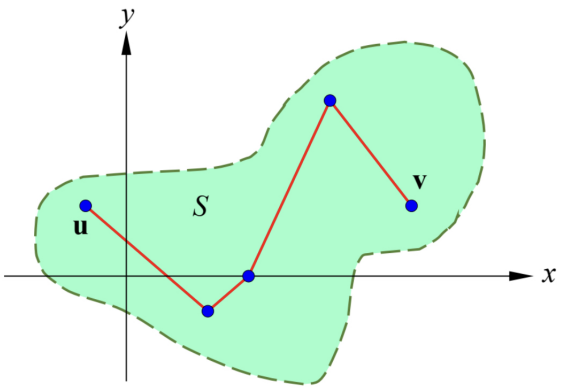


Figure 3.9: A polygonal path.

If $\boldsymbol{\gamma}_1 : [a, c] \to A$ is a polygonal path in A that joins $\mathbf{u}$ to $\mathbf{w}$, $\boldsymbol{\gamma}_2 : [c, b] \to B$ is a polygonal path in B that joins $\mathbf{w}$ to $\mathbf{v}$, then the path $\boldsymbol{\gamma} : [a, b] \to A \cup B$,

$$\boldsymbol{\gamma}(t) = \begin{cases} \boldsymbol{\gamma}_1(t), & \text{if } a \leq t \leq c, \\ \boldsymbol{\gamma}_2(t), & \text{if } c \leq t \leq b, \end{cases}$$

is a polygonal path in $A \cup B$ that joins $\mathbf{u}$ to $\mathbf{v}$. Using this, we can prove the following useful theorem.

Theorem 3.16

Let S be a connected subset of $\mathbb{R}^n$. If S is an open set, then any two points in S can be joined by a polygonal path in S. In particular, S is path connected.

Proof

We use proof by contradiction. Supposed that S is open but there are two points $\mathbf{u}$ and $\mathbf{v}$ in S that cannot be joined by a polygonal path in S. Consider the sets

$$U = \{\mathbf{x} \in S \,|\, \text{there is a polygonal path in } S \text{ that joins } \mathbf{u} \text{ to } \mathbf{x}\},$$

$$V = \{\mathbf{x} \in S \,|\, \text{there is no polygonal path in } S \text{ that joins } \mathbf{u} \text{ to } \mathbf{x}\}.$$

Obviously $\mathbf{u}$ is in U and $\mathbf{v}$ is in V, and $S = U \cup V$. We claim that both U and V are open sets.

If $\mathbf{x}$ is in the open set S, there is an $r > 0$ such that $B(\mathbf{x}, r) \subset S$. Since $B(\mathbf{x}, r)$ is convex, any point $\mathbf{w}$ in $B(\mathbf{x}, r)$ can be joined by a polygonal path in $B(\mathbf{x}, r)$ to $\mathbf{x}$. Hence, if $\mathbf{x}$ is in U, $\mathbf{w}$ is in U. If $\mathbf{x}$ is in V, $\mathbf{w}$ is in V. This shows that if $\mathbf{x}$ is in U, then $B(\mathbf{x}, r) \subset U$. If $\mathbf{x}$ is in V, then $B(\mathbf{x}, r) \subset V$. Hence, U and V are open sets.
Since U and V are nonempty open sets and $S = U \cup V$, they form a separation of S. This contradicts to S is connected. Hence, any two points in S can be joined by a polygonal path in S.

Exercises 3.2

Question 1

Determine whether the set

$$A = \{(x, y) \,|\, y = 0\} \cup \left\{ (x, y) \;\middle|\; x > 0, y = \frac{2}{x} \right\}$$

is connected.

Question 2

Let $\mathfrak{D}$ be a connected subset of $\mathbb{R}^n$, and let $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ be a function defined on $\mathfrak{D}$. If $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is continuous, show that the graph of $\mathbf{F}$,

$$G_{\mathbf{F}} = \{(\mathbf{x}, \mathbf{y}) \,|\, \mathbf{x} \in \mathfrak{D}, \mathbf{y} = \mathbf{F}(\mathbf{x})\}$$

is also connected.

Question 3

Determine whether the set

$$A = \{(x, y) \,|\, 0 \leq x < 1, -1 < y \leq 1\} \cup \{(1, 0), (1, 1)\}$$

is connected.

Question 4

Assume that A is a connected subset of $\mathbb{R}^3$ that contains the points $\mathbf{u} = (0, 2, 0)$ and $\mathbf{v} = (2, -6, 3)$.

(a) Show that there is a point $\mathbf{x} = (x, y, z)$ in A that lies in the plane $y = 0$.

(b) Show that there exists a point $\mathbf{x} = (x, y, z)$ in A that lies on the sphere $x^2 + y^2 + z^2 = 25$.

Question 5

Let A and B be connected subsets of $\mathbb{R}^n$. If $A \cap B$ is nonempty, show that $S = A \cup B$ is connected.

3.3 Sequential Compactness and Compactness

In volume I, we have seen that sequential compactness plays important role in extreme value theorem. In this section, we extend the definition of sequential compactness to subsets of $\mathbb{R}^n$. We will also consider another concept called compactness.

Let us start with the definition of bounded sets.

Definition 3.9 Bounded Sets

Let S be a subset of $\mathbb{R}^n$. We say that S is bounded if there exists a positive number M such that

$$\|\mathbf{x}\| \leq M \qquad \text{for all } \mathbf{x} \in S.$$

Remark 3.3

Let S be a subset of $\mathbb{R}^n$. If S is bounded and S' is a subset of S, then it is obvious that S' is also bounded.

Example 3.15

Show that a ball $B(\mathbf{x}_0, r)$ in $\mathbb{R}^n$ is bounded.

Solution

Given $\mathbf{x} \in B(\mathbf{x}_0, r)$, $\|\mathbf{x} - \mathbf{x}_0\| < r$. Thus,

$$\|\mathbf{x}\| \leq \|\mathbf{x}_0\| + \|\mathbf{x} - \mathbf{x}_0\| < \|\mathbf{x}_0\| + r.$$

Since $M = \|\mathbf{x}_0\| + r$ is a constant independent of the points in the ball $B(\mathbf{x}_0, r)$, the ball $B(\mathbf{x}_0, r)$ is bounded.

Notice that if $\mathbf{x}_1$ and $\mathbf{x}_2$ are points in $\mathbb{R}^n$, and S is a set in $\mathbb{R}^n$ such that

$$\|\mathbf{x} - \mathbf{x}_1\| < r_1 \qquad \text{for all } \mathbf{x} \in S,$$

then

$$\|\mathbf{x} - \mathbf{x}_2\| < r_1 + \|\mathbf{x}_2 - \mathbf{x}_1\| \qquad \text{for all } \mathbf{x} \in S.$$

Thus, we have the following.

Proposition 3.17

Let S be a subset in $\mathbb{R}^n$. The following are equivalent.

(a) S is bounded.

(b) There is a point $\mathbf{x}_0$ in $\mathbb{R}^n$ and a positive constant M such that

$$\|\mathbf{x} - \mathbf{x}_0\| \leq M \qquad \text{for all } \mathbf{x} \in S.$$

(c) For any $\mathbf{x}_0$ in $\mathbb{R}^n$, there is a positive constant M such that

$$\|\mathbf{x} - \mathbf{x}_0\| \leq M \qquad \text{for all } \mathbf{x} \in S.$$

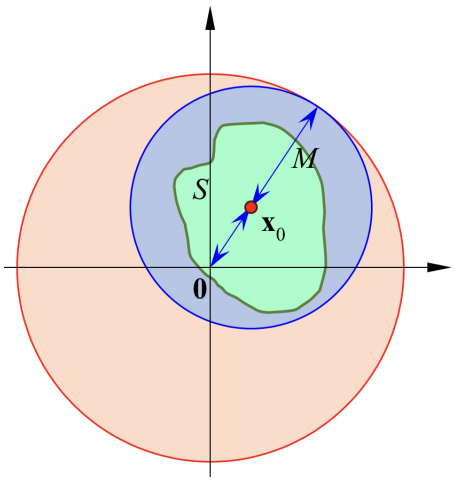


Figure 3.10: The set S is bounded.

We say that a sequence $\{\mathbf{x}_k\}$ is bounded if the set $\{\mathbf{x}_k \,|\, k \in \mathbb{Z}^+\}$ is bounded. The following is a standard theorem about convergent sequences.

Proposition 3.18

If $\{\mathbf{x}_k\}$ is a sequence in $\mathbb{R}^n$ that is convergent, then it is bounded.

Proof

Assume that the sequence $\{\mathbf{x}_k\}$ converges to the point $\mathbf{x}_0$. Then there is a positive integer K such that

$$\|\mathbf{x}_k - \mathbf{x}_0\| < 1 \qquad \text{for all } k \geq K.$$

Let

$$M = \max\{\|\mathbf{x}_k - \mathbf{x}_0\| \,|\, 1 \leq k \leq K-1\} + 1.$$

Then M is finite and

$$\|\mathbf{x}_k - \mathbf{x}_0\| \leq M \qquad \text{for all } k \in \mathbb{Z}^+.$$

Hence, the sequence $\{x_k\}$ is bounded.

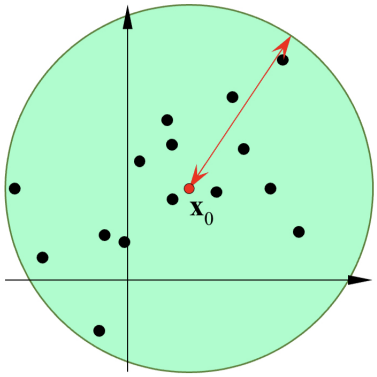


Figure 3.11: A convergent sequence is bounded.

Let us now define the *diameter* of a bounded set. If S is a subset of $\mathbb{R}^n$ that is bounded, there is a positive number M such that

$$\|\mathbf{x}\| \leq M \qquad \text{for all } \mathbf{x} \in S.$$

It follows from triangle inequality that for any $\mathbf{u}$ and $\mathbf{v}$ in S,

$$\|\mathbf{u} - \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\| \leq 2M.$$

Thus, the set

$$D_S = \{d(\mathbf{u}, \mathbf{v}) \,|\, \mathbf{u}, \mathbf{v} \in S\} = \{\|\mathbf{u} - \mathbf{v}\| \,|\, \mathbf{u}, \mathbf{v} \in S\} \tag{3.1}$$

is a set of nonnegative real numbers that is bounded above. In fact, for any subset S of $\mathbb{R}^n$, one can define the set of real numbers D_S by (3.1). Then S is a bounded set if and only if the set D_S is bounded above.

Definition 3.10 Diameter of a Bounded Set

Let S be a bounded subset of $\mathbb{R}^n$. The diameter of S, denoted by diam S, is defined as

$$\text{diam}\, S = \sup\{d(\mathbf{u}, \mathbf{v}) \mid \mathbf{u}, \mathbf{v} \in S\} = \sup\{\|\mathbf{u} - \mathbf{v}\| \mid \mathbf{u}, \mathbf{v} \in S\}.$$

Example 3.16

Consider the rectangle $R = [a_1, b_1] \times \cdots \times [a_n, b_n]$. If $\mathbf{u}$ and $\mathbf{v}$ are two points in R, for each $1 \leq i \leq n$, $u_i, v_i \in [a_i, b_i]$. Thus,

$$|u_i - v_i| \leq b_i - a_i.$$

It follows that

$$\|\mathbf{u} - \mathbf{v}\| \leq \sqrt{(b_1 - a_1)^2 + \cdots + (b_n - a_n)^2}.$$

If $\mathbf{u}_0 = \mathbf{a} = (a_1, \ldots, a_n)$ and $\mathbf{v}_0 = \mathbf{b} = (b_1, \ldots, b_n)$, then $\mathbf{u}_0$ and $\mathbf{v}_0$ are in R, and

$$\|\mathbf{u}_0 - \mathbf{v}_0\| = \sqrt{(b_1 - a_1)^2 + \cdots + (b_n - a_n)^2}.$$

This shows that the diameter of R is

$$\text{diam}\, R = \|\mathbf{b} - \mathbf{a}\| = \sqrt{(b_1 - a_1)^2 + \cdots + (b_n - a_n)^2}.$$

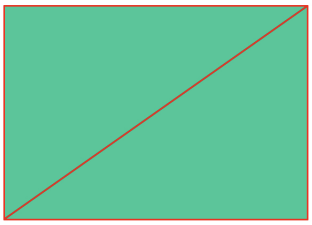


Figure 3.12: The diameter of a rectangle.

Intuitively, the diameter of the open rectangle $U = (a_1, b_1) \times \cdots \times (a_n, b_n)$ is also equal to

$$d = \sqrt{(b_1 - a_1)^2 + \cdots + (b_n - a_n)^2}.$$

However, the points $\mathbf{a} = (a_1, \ldots, a_n)$ and $\mathbf{b} = (b_1, \ldots, b_n)$ are not in U. There does not exist two points in U whose distance is d, but there are sequences of points $\{\mathbf{u}_k\}$ and $\{\mathbf{v}_k\}$ such that their distances $\{\|\mathbf{u}_k - \mathbf{v}_k\|\}$ approach d as $k \to \infty$. We will formulate this as a more general theorem.

Theorem 3.19

Let S be a subset of $\mathbb{R}^n$. If S is bounded, then its closure $\overline{S}$ is also bounded. Moreover, $\operatorname{diam} \overline{S} = \operatorname{diam} S$.

Proof

If $\mathbf{u}$ and $\mathbf{v}$ are two points in $\overline{S}$, there exist sequences $\{\mathbf{u}_k\}$ and $\{\mathbf{v}_k\}$ in S that converge respectively to $\mathbf{u}$ and $\mathbf{v}$. Then

$$d(\mathbf{u}, \mathbf{v}) = \lim_{k \to \infty} d(\mathbf{u}_k, \mathbf{v}_k). \tag{3.2}$$

For each $k \in \mathbb{Z}^+$, since $\mathbf{u}_k$ and $\mathbf{v}_k$ are in S,

$$d(\mathbf{u}_k, \mathbf{v}_k) \leq \operatorname{diam} S.$$

Eq. (3.2) implies that

$$d(\mathbf{u}, \mathbf{v}) \leq \operatorname{diam} S.$$

Since this is true for any $\mathbf{u}$ and $\mathbf{v}$ in $\overline{S}$, $\overline{S}$ is bounded and

$$\operatorname{diam} \overline{S} \leq \operatorname{diam} S.$$

Since $S \subset \overline{S}$, we also have $\operatorname{diam} S \leq \operatorname{diam} \overline{S}$. We conclude that $\operatorname{diam} \overline{S} = \operatorname{diam} S$.

The following example justifies that the diameter of a ball of radius r is indeed $2r$.

Example 3.17

Find the diameter of the open ball $B(\mathbf{x}_0, r)$ in $\mathbb{R}^n$.

Solution

By Theorem 3.19, the diameter of the open ball $B(\mathbf{x}_0, r)$ is the same as the diameter of its closure, the closed ball $CB(\mathbf{x}_0, r)$. Given $\mathbf{u}$ and $\mathbf{v}$ in $CB(\mathbf{x}_0, r)$, $\|\mathbf{u} - \mathbf{x}_0\| \leq r$ and $\|\mathbf{v} - \mathbf{x}_0\| \leq r$. Therefore,

$$\|\mathbf{u} - \mathbf{v}\| \leq \|\mathbf{u} - \mathbf{x}_0\| + \|\mathbf{v} - \mathbf{x}_0\| \leq 2r.$$

This shows that $\operatorname{diam} CB(\mathbf{x}_0, r) \leq 2r$. The points $\mathbf{u}_0 = \mathbf{x}_0 + r\mathbf{e}_1$ and $\mathbf{v}_0 = \mathbf{x}_0 - r\mathbf{e}_1$ are in the closed ball $CB(\mathbf{x}_0, r)$. Since

$$\|\mathbf{u}_0 - \mathbf{v}_0\| = \|2r\mathbf{e}_1\| = 2r,$$

$\operatorname{diam} CB(\mathbf{x}_0, r) \geq 2r$. Therefore, the diameter of the closed ball $CB(\mathbf{x}_0, r)$ is exactly $2r$. By Theorem 3.19, the diameter of the open ball $B(\mathbf{x}_0, r)$ is also $2r$.

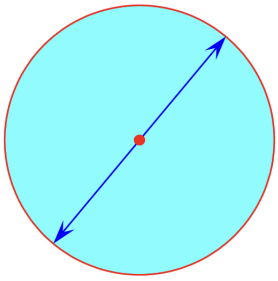


Figure 3.13: The diameter of a ball.

In volume I, we have shown that a bounded sequence in $\mathbb{R}$ has a convergent subsequence. This is achieved by using the monotone convergence theorem, which says that a bounded monotone sequence in $\mathbb{R}$ is convergent. For points in $\mathbb{R}^n$ with $n \geq 2$, we cannot apply monotone convergence theorem, as we cannot define a simple order on the points in $\mathbb{R}^n$ when $n \geq 2$. Nevertheless, we can use the result of $n = 1$ and the componentwise convergence theorem to show that a bounded sequence in $\mathbb{R}^n$ has a convergent subsequence.

Theorem 3.20

Let $\{\mathbf{u}_k\}$ be a sequence in $\mathbb{R}^n$. If $\{\mathbf{u}_k\}$ is bounded, then there is a subsequence that is convergent.

Sketch of Proof

The $n = 1$ case is already established in volume I. Here we prove the $n = 2$ case. The $n \geq 3$ case can be proved by induction using the same reasoning. For $k \in \mathbb{Z}^+$, let $\mathbf{u}_k = (x_k, y_k)$. Since

$$|x_k| \leq \|\mathbf{u}_k\| \quad \text{and} \quad |y_k| \leq \|\mathbf{u}_k\|,$$

the sequences $\{x_k\}$ and $\{y_k\}$ are bounded sequences. Thus, there is a subsequence $\{x_{k_j}\}_{j=1}^{\infty}$ of $\{x_k\}_{k=1}^{\infty}$ that converges to a point x_0 in $\mathbb{R}$. Consider the subsequence $\{y_{k_j}\}_{j=1}^{\infty}$ of the sequence $\{y_k\}_{k=1}^{\infty}$. It is also bounded. Hence, there is a subsequence $\{y_{k_{j_l}}\}_{l=1}^{\infty}$ that converges to a point y_0 in $\mathbb{R}$. Notice that the subsequence $\{x_{k_{j_l}}\}_{l=1}^{\infty}$ of $\{x_k\}_{k=1}^{\infty}$ is also a subsequence of $\{x_{k_j}\}_{j=1}^{\infty}$. Hence, it also converges to x_0. By componentwise convergence theorem, $\{\mathbf{u}_{k_{j_l}}\}_{l=1}^{\infty}$ is a subsequence of $\{\mathbf{u}_k\}_{k=1}^{\infty}$ that converges to (x_0, y_0). This proves the theorem when $n = 2$.

Now we study the concept of sequential compactness. It is the same as the $n = 1$ case.

Definition 3.11 Sequentially Compact

Let S be a subset of $\mathbb{R}^n$. We say that S is sequentially compact provided that every sequence in S has a subsequence that converges to a point in S.

In volume I, we proved the Bolzano-Weierstrass theorem, which says that a subset of $\mathbb{R}$ is sequentially compact if and only if it is closed and bounded. In fact, the same is true for the $n \geq 2$ case. Let us first look at some examples.

Example 3.18

Show that the set $A = \{(x, y) \,|\, x^2 + y^2 < 1\}$ is not sequentially compact.

Solution

For $k \in \mathbb{Z}^+$, let

$$\mathbf{u}_k = \left(\frac{k}{k+1}, 0 \right).$$

Then $\{\mathbf{u}_k\}$ is a sequence in A that converges to the point $\mathbf{u}_0 = (1, 0)$ that is not in A. Thus, every subsequence of $\{\mathbf{u}_k\}$ converges to the point $\mathbf{u}_0$, which is not in A. This means the sequence $\{\mathbf{u}_k\}$ in A does not have a subsequence that converges to a point in A. Hence, A is not sequentially compact.

Note that the set A in Example 3.18 is not closed.

Example 3.19

Show that the set $C = \{(x, y) \,|\, 1 \leq x \leq 3, y \geq 0\}$ is not sequentially compact.

Solution

For $k \in \mathbb{Z}^+$, let $\mathbf{u}_k = (2, k)$. Then $\{\mathbf{u}_k\}$ is a sequence in C. If $\{\mathbf{u}_{k_j}\}_{j=1}^{\infty}$ is a subsequence of $\{\mathbf{u}_k\}$, then $k_1, k_2, k_3, \ldots$ is a strictly increasing sequence of positive integers. Therefore $k_j \geq j$ for all $j \in \mathbb{Z}^+$. It follows that

$$\|\mathbf{u}_{k_j}\| = \|(2, k_j)\| \geq k_j \geq j \qquad \text{for all } j \in \mathbb{Z}^+.$$

Hence, the subsequence $\{\mathbf{u}_{k_j}\}$ is not bounded. Therefore, it is not convergent. This means that the sequence $\{\mathbf{u}_k\}$ in C does not have a convergent subsequence. Therefore, C is not sequentially compact.

Note that the set C in Example 3.19 is not bounded.

Now we prove the main theorem.

Theorem 3.21 Bolzano-Weierstrass Theorem

Let S be a subset of $\mathbb{R}^n$. The following are equivalent.

(a) S is closed and bounded.

(b) S is sequentially compact.

Proof

First assume that S is closed and bounded. Let $\{\mathbf{x}_k\}$ be a sequence in S. Then $\{\mathbf{x}_k\}$ is also bounded. By Theorem 3.20, there is subsequence $\{\mathbf{x}_{k_j}\}$ that converges to some $\mathbf{x}_0$. Since S is closed, we must have $\mathbf{x}_0$ is in S. This proves that every sequence in S has a subsequence that converges to a point in S. Hence, S is sequentially compact. This completes the proof of (a) implies (b).

To prove that (b) implies (a), it suffices to show that if S is not closed or S is not bounded, then S is not sequentially compact.

If S is not closed, there is a sequence $\{\mathbf{x}_k\}$ in S that converges to a point $\mathbf{x}_0$, but $\mathbf{x}_0$ is not in S. Then every subsequence of $\{\mathbf{x}_k\}$ converges to the point $\mathbf{x}_0$, which is not in S. Thus, $\{\mathbf{x}_k\}$ is a sequence in S that does not have any subsequence that converges to a point in S. This shows that S is not sequentially compact.

If S is not bounded, for each positive integer k, there is a point $\mathbf{x}_k$ in S such that $\|\mathbf{x}_k\| \geq k$. If $\{\mathbf{x}_{k_j}\}_{j=1}^{\infty}$ is a subsequence of $\{\mathbf{x}_k\}$, then $k_1, k_2, k_3, \ldots$ is a strictly increasing sequence of positive integers. Therefore $k_j \geq j$ for all $j \in \mathbb{Z}^+$. It follows that $\|\mathbf{x}_{k_j}\| \geq k_j \geq j$ for al $j \in \mathbb{Z}^+$. Hence, the subsequence $\{\mathbf{x}_{k_j}\}$ is not bounded. Therefore, it is not convergent. This means that the sequence $\{\mathbf{x}_k\}$ in S does not have a convergent subsequence. Therefore, S is not sequentially compact.

Corollary 3.22

A closed rectangle $R = [a_1, b_1] \times \cdots \times [a_n, b_n]$ in $\mathbb{R}^n$ is sequentially compact.

Proof

We have shown in Chapter 1 that R is closed. Example 3.16 shows that R is bounded. Thus, R is sequentially compact.

An interesting consequence of Theorem 3.19 is the following.

Corollary 3.23

If S be a bounded subset of $\mathbb{R}^n$, then its closure $\overline{S}$ is sequentially compact.

Example 3.20

Determine whether the following subsets of $\mathbb{R}^3$ is sequentially compact.

(a) $A = \{(x, y, z) \,|\, xyz = 1\}$.

(b) $B = \{(x, y, z) \,|\, x^2 + 4y^2 + 9z^2 \leq 36\}$.

(c) $C = \{(x, y, z) \,|\, 1 \leq x \leq 2, 1 \leq y \leq 3, 0 < xyz \leq 4\}$.

Solution

(a) For any $k \in \mathbb{Z}^+$, let

$$\mathbf{u}_k = \left(k, \frac{1}{k}, 1\right).$$

Then $\{\mathbf{u}_k\}$ is a sequence in A, and $\|\mathbf{u}_k\| \geq k$. Therefore, A is not bounded. Hence, A is not sequentially compact.

(b) For any $\mathbf{u} = (x, y, z) \in B$,

$$\|\mathbf{u}\|^2 = x^2 + y^2 + z^2 \leq x^2 + 4y^2 + 9z^2 \leq 36.$$

Hence, B is bounded. The function $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x, y, z) = x^2 + 4y^2 + 9z^2$ is a polynomial. Hence, it is continuous. Since the set $I = (-\infty, 36]$ is closed in $\mathbb{R}$, and $B = f^{-1}(I)$, B is closed in $\mathbb{R}^3$. Since B is closed and bounded, it is sequentially compact.

(c) For any $k \in \mathbb{Z}^+$, let

$$\mathbf{u}_k = \left(1, 1, \frac{1}{k}\right).$$

Then $\{\mathbf{u}_k\}$ is a sequence of points in C that converges to the point $\mathbf{u}_0 = (1, 1, 0)$, which is not in C. Thus, C is not closed, and so C is not sequentially compact.

The following theorem asserts that continuous functions preserve sequential compctness.

Theorem 3.24

Let $\mathfrak{D}$ be a sequentially compact subset of $\mathbb{R}^n$. If the function $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is continuous, then $\mathbf{F}(\mathfrak{D})$ is a sequentially compact subset of $\mathbb{R}^m$.

The proof of this theorem is identical to the $n = 1$ case.

Proof

Let $\{\mathbf{y}_k\}$ be a sequence in $\mathbf{F}(\mathfrak{D})$. For each $k \in \mathbb{Z}^+$, there exists $\mathbf{x}_k \in \mathfrak{D}$ such that $\mathbf{F}(\mathbf{x}_k) = \mathbf{y}_k$. Since $\mathfrak{D}$ is sequentially compact, there is a subsequence $\{\mathbf{x}_{k_j}\}$ of $\{\mathbf{x}_k\}$ that converges to a point $\mathbf{x}_0$ in $\mathfrak{D}$. Since $\mathbf{F}$ is continuous, the sequence $\{\mathbf{F}(\mathbf{x}_{k_j})\}$ converges to $\mathbf{F}(\mathbf{x}_0)$. Note that $\mathbf{F}(\mathbf{x}_0)$ is in $\mathbf{F}(\mathfrak{D})$. In other words, $\{\mathbf{y}_{k_j}\}$ is a subsequence of the sequence $\{\mathbf{y}_k\}$ that converges to $\mathbf{F}(\mathbf{x}_0)$ in $\mathbf{F}(\mathfrak{D})$. This shows that every sequence in $\mathbf{F}(\mathfrak{D})$ has a subsequence that converges to a point in $\mathbf{F}(\mathfrak{D})$. Thus, $\mathbf{F}(\mathfrak{D})$ is a sequentially compact subset of $\mathbb{R}^m$.

We are going to discuss important consequences of Theorem 3.24 in the coming section. For the rest of this section, we introduce the concept of compactness, which plays a central role in modern analysis. We start with the definition of an open covering.

Definition 3.12 Open Covering

Let S be a subset of $\mathbb{R}^n$, and let $\mathscr{A} = \{U_\alpha \mid \alpha \in J\}$ be a collection of open sets in $\mathbb{R}^n$ indexed by the set J. We say that $\mathscr{A}$ is an open covering of S provided that

$$S \subset \bigcup_{\alpha \in J} U_\alpha.$$

Example 3.21

For each $k \in \mathbb{Z}^+$, let $U_k = (1/k, 1)$. Then U_k is an open set in $\mathbb{R}$ and

$$\bigcup_{k=1}^{\infty} U_k = (0, 1).$$

Hence, $\mathscr{A} = \{U_k \mid k \in \mathbb{Z}^+\}$ is an open covering of the set $S = (0, 1)$.

Remark 3.4

If $\mathscr{A} = \{U_\alpha \mid \alpha \in J\}$ is an open covering of S and S' is a subset of S, then $\mathscr{A} = \{U_\alpha \mid \alpha \in J\}$ is also an open covering of S'.

Example 3.22

For each $k \in \mathbb{Z}^+$, let $U_k = B(\mathbf{0}, k)$ be the ball in $\mathbb{R}^n$ centered at the origin and having radius k. Then

$$\bigcup_{k=1}^{\infty} U_k = \mathbb{R}^n.$$

Thus, $\mathscr{A} = \{U_k \mid k \in \mathbb{Z}^+\}$ is an open covering of any subset S of $\mathbb{R}^n$.

Definition 3.13 Subcover

Let S be a subset of $\mathbb{R}^n$, and let $\mathscr{A} = \{U_\alpha \mid \alpha \in J\}$ be an open covering of S. A subcover is a subcollection of $\mathscr{A}$ which is also a covering of S. A *finite subcover* is a subcover that contains only finitely many elements.

Example 3.23

For each $k \in \mathbb{Z}$, let $U_k = (k, k+2)$. Then $\bigcup_{k=-\infty}^{\infty} U_k = \mathbb{R}$. Thus, $\mathscr{A} = \{U_k \,|\, k \in \mathbb{Z}\}$ is an open covering of the set $S = [-3, 4)$. There is a finite subcover of S given by

$$\mathscr{B} = \{U_{-4}, U_{-3}, U_{-2}, U_{-1}, U_0, U_1, U_2\}.$$

Definition 3.14 Compact Sets

Let S be a subset of $\mathbb{R}^n$. We say that S is compact provided that every open covering of S has a finite subcover. Namely, if $\mathscr{A} = \{U_\alpha \,|\, \alpha \in J\}$ is an open covering of S, then there exist $\alpha_1, \ldots, \alpha_k \in J$ such that

$$S \subset \bigcup_{j=1}^{k} U_{\alpha_j}.$$

Example 3.24

The subset $S = (0, 1)$ of $\mathbb{R}$ is not compact. For $k \in \mathbb{Z}^+$, let $U_k = (1/k, 1)$. Example 3.21 says that $\mathscr{A} = \{U_k \,|\, k \in \mathbb{Z}^+\}$ is an open covering of the set S. We claim that there is no finite subcollection of $\mathscr{A}$ that covers S.

Assume to the contrary that there exists a finite subcollection of $\mathscr{A}$ that covers S. Then there are positive integers $k_1, \ldots, k_m$ such that

$$(0, 1) \subset \bigcup_{j=1}^{m} U_{k_j} = \bigcup_{j=1}^{m} \left(\frac{1}{k_j}, 1\right).$$

Notice that if $k_i \leq k_j$, then $U_{k_i} \subset U_{k_j}$. Thus, if $K = \max\{k_1, \ldots, k_m\}$, then

$$\bigcup_{j=1}^{m} U_{k_j} = U_K = \left(\frac{1}{K}, 1\right),$$

and so $S = (0, 1)$ is not contained in U_K. This gives a contradiction. Hence, S is not compact.

Example 3.25

As a subset of itself, $\mathbb{R}^n$ is not compact. For $k \in \mathbb{Z}^+$, let $U_k = B(\mathbf{0}, k)$ be the ball in $\mathbb{R}^n$ centered at the origin and having radius k. Example 3.22 says that $\mathscr{A} = \{U_k \,|\, k \in \mathbb{Z}^+\}$ is an open covering of $\mathbb{R}^n$. We claim that there is no finite subcover.

Assume to the contrary that there is a finite subcover. Then there exist positive integers $k_1, \ldots, k_m$ such that

$$\mathbb{R}^n = \bigcup_{j=1}^{m} U_k.$$

Notice that if $k_i \leq k_j$, then $U_{k_i} \subset U_{k_j}$. Thus, if $K = \max\{k_1, \ldots, k_m\}$, then

$$\bigcup_{j=1}^{m} U_{k_j} = U_K = B(\mathbf{0}, K).$$

Obviously, $B(\mathbf{0}, K)$ is not equal to $\mathbb{R}^n$. This gives a contradiction. Hence, $\mathbb{R}^n$ is not compact.

Our goal is to prove the Heine-Borel theorem, which says that a subset of $\mathbb{R}^n$ is compact if and only if it is closed and bounded. We first prove the easier direction.

Theorem 3.25

Let S be a subset of $\mathbb{R}^n$. If S is compact, then it is closed and bounded.

Proof

We show that if S is compact, then it is bounded; and if S is compact, then it is closed.

First we prove that if S is compact, then it is bounded. For $k \in \mathbb{Z}^+$, let $U_k = B(\mathbf{0}, k)$ be the ball in $\mathbb{R}^n$ centered at the origin and having radius k. Example 3.22 says that $\mathscr{A} = \{U_k \,|\, k \in \mathbb{Z}^+\}$ is an open covering of S. Since S is compact, there exist positive integers $k_1, \ldots, k_m$ such that

$$S \subset \bigcup_{j=1}^{m} U_{k_j} = U_K = B(\mathbf{0}, K),$$

where $K = \max\{k_1, \ldots, k_m\}$. This shows that

$$\|\mathbf{x}\| \leq K \qquad \text{for all } \mathbf{x} \in S.$$

Hence, S is bounded.

Now we prove that if S is compact, then it is closed. For this, it suffices to show that $\overline{S} \subset S$, or equivalently, any point that is not in S is not in $\overline{S}$. Assume that $\mathbf{x}_0$ is not in S. For each $k \in \mathbb{Z}^+$, let

$$V_k = \operatorname{ext} B(\mathbf{x}_0, 1/k) = \left\{ \mathbf{x} \in \mathbb{R}^n \;\middle|\; \|\mathbf{x} - \mathbf{x}_0\| > \frac{1}{k} \right\}.$$

Then V_k is open in $\mathbb{R}^n$. If $\mathbf{x}$ is a point in $\mathbb{R}^n$ and $\mathbf{x} \neq \mathbf{x}_0$, then $r = \|\mathbf{x} - \mathbf{x}_0\| > 0$. There is a $k \in \mathbb{Z}^+$ such that $1/k < r$. Then $\mathbf{x}$ is in V_k. This shows that

$$\bigcup_{k=1}^{\infty} V_k = \mathbb{R}^n \setminus \{\mathbf{x}_0\}.$$

Therefore, $\mathscr{A} = \{V_k \,|\, k \in \mathbb{Z}^+\}$ is an open covering of S. Since S is compact, there is a finite subcover. Namely, there exist positive integers $k_1, \ldots, k_m$ such that

$$S \subset \bigcup_{j=1}^{m} V_{k_j} = V_K,$$

where $K = \max\{k_1, \ldots, k_m\}$. Since $B(\mathbf{x}_0, 1/K)$ is disjoint from V_K, it does not contain any point of S. This shows that $\mathbf{x}_0$ is not in $\overline{S}$, and thus the proof is completed.

Example 3.26

The set

$$A = \{(x, y, z) \,|\, xyz = 1\}$$

in Example 3.20 is not compact because it is not bounded. The set

$$C = \{(x, y, z) \,|\, 1 \leq x \leq 2, 1 \leq y \leq 3, 0 < xyz \leq 4\}$$

is not compact because it is not closed.

We are now left to show that a closed and bounded subset of $\mathbb{R}^n$ is compact. We start by proving a special case.

Theorem 3.26

A closed rectangle $R = [a_1, b_1] \times \cdots \times [a_n, b_n]$ in $\mathbb{R}^n$ is compact.

Proof

We will prove by contradiction. Assume that R is not compact, and we show that this will lead to a contradiction. The idea is to use the bisection method.

If R is not compact, there is an open covering $\mathscr{A} = \{U_\alpha \,|\, \alpha \in J\}$ of R which does not have a finite subcover.

Let $R_1 = R$, and let $d_1 = \operatorname{diam} R_1$. For $1 \le i \le n$, let $a_{i,1} = a_i$ and $b_{i,1} = b_i$, and let $m_{i,1}$ to be the midpoint of the interval $[a_{i,1}, b_{i,1}]$. The hyperplanes $x_i = m_{i,1}$, $1 \le i \le n$, divides the rectangle R_1 into 2^n subrectangles. Notice that $\mathscr{A}$ is also an open covering of each of these subrectangles. If each of these subrectangles can be covered by a finite subcollection of open sets in $\mathscr{A}$, then R also can be covered by a finite subcollection of open sets in $\mathscr{A}$. Since we assume R cannot be covered by any finite subcollection of open sets in $\mathscr{A}$, there is at least one of the 2^n subrectangles which cannot be covered by any finite subcollection of open sets in $\mathscr{A}$. Choose one of these, and denote it by R_2.

Define $a_{i,2}, b_{i,2}$ for $1 \le i \le n$ so that

$$R_2 = [a_{1,2}, b_{1,2}] \times \cdots \times [a_{n,2}, b_{n,2}].$$

Note that

$$b_{i,2} - a_{i,2} = \frac{b_{i,1} - a_{i,1}}{2} \qquad \text{for } 1 \le i \le n.$$

Therefore, $d_2 = \operatorname{diam} R_2 = d_1/2$.

We continue this bisection process to obtain the rectangles $R_1, R_2, \cdots$, so that $R_{k+1} \subset R_k$ for all $k \in \mathbb{Z}^+$, and R_k cannot be covered by any finite subcollections of $\mathscr{A}$.

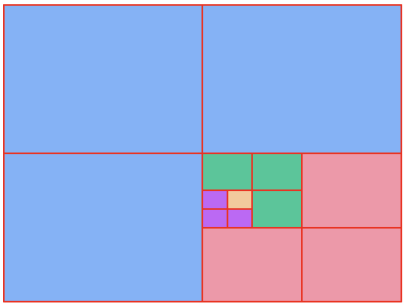


Figure 3.14: Bisection method.

Define $a_{i,k}, b_{i,k}$ for $1 \le i \le n$ so that

$$R_k = [a_{1,k}, b_{1,k}] \times \cdots \times [a_{n,k}, b_{n,k}].$$

Then for all $k \in \mathbb{Z}^+$,

$$b_{i,k+1} - a_{i,k+1} = \frac{b_{i,k} - a_{i,k}}{2} \qquad \text{for } 1 \le i \le n.$$

It follows that $d_{k+1} = \operatorname{diam} R_{k+1} = d_k/2$.
For any $1 \le i \le n$, $\{a_{i,k}\}_{k=1}^{\infty}$ is an increasing sequence that is bounded above by b_i, and $\{b_{i,k}\}_{k=1}^{\infty}$ is a decreasing sequence that is bounded below by a_i. By monotone convergence theorem, the sequence $\{a_{i,k}\}_{k=1}^{\infty}$ converges to $a_{i,0} = \sup_{k\in\mathbb{Z}^+} a_{i,k}$; while the sequence $\{b_{i,k}\}_{k=1}^{\infty}$ converges to $b_{i,0} = \inf_{k\in\mathbb{Z}^+} b_{i,k}$. Since

$$b_{i,k} - a_{i,k} = \frac{b_i - a_i}{2^{k-1}} \qquad \text{for all } k \in \mathbb{Z}^+,$$

we find that $a_{i,0} = b_{i,0}$. Let $c_i = a_{i,0} = b_{i,0}$. Then $a_{i,k} \le c_i \le b_{i,k}$ for all $1 \le i \le n$ and all $k \in \mathbb{Z}^+$. Thus, $\mathbf{c} = (c_1, \ldots, c_n)$ is a point in R_k for all $k \in \mathbb{Z}^+$. By assumption that $\mathscr{A}$ is an open covering of $R = R_1$, there exists $\beta \in J$ such that $\mathbf{c} \in U_\beta$. Since U_β is an open set, there is an $r > 0$ such that $B(\mathbf{c}, r) \subset U_\beta$. Since

$$d_k = \operatorname{diam} R_k = \frac{d_1}{2^{k-1}} \qquad \text{for all } k \in \mathbb{Z}^+,$$

we find that $\lim_{k\to\infty} d_k = 0$. Hence, there is a positive integer K such that $d_K < r$. If $\mathbf{x} \in R_K$, then

$$\|\mathbf{x} - \mathbf{c}\| \leq \text{diam}\, R_K = d_K < r.$$

This implies that $\mathbf{x}$ is in $B(\mathbf{c}, r)$. Thus, we have shown that $R_K \subset B(\mathbf{c}, r)$. Therefore, R_K is contained in the single element U_β of $\mathscr{A}$, which contradicts to R_K cannot be covered by any finite subcollection of open sets in $\mathscr{A}$.
We conclude that R must be compact.

Now we can prove the Heine-Borel theorem.

Theorem 3.27 Heine-Borel Theorem

Let S be a subset of $\mathbb{R}^n$. Then S is compact if and only if it is closed and bounded.

Proof

We have shown in Theorem 3.25 that if S is compact, then it must be closed and bounded.
Now assume that S is closed and bounded, and let $\mathscr{A} = \{U_\alpha \,|\, \alpha \in J\}$ be an open covering of S. Since S is bounded, there exists a positive number M such that

$$\|\mathbf{x}\| \leq M \qquad \text{for all } \mathbf{x} \in S.$$

Thus, if $\mathbf{x} = (x_1, \ldots, x_n)$ is in S, then for all $1 \leq i \leq n$, $|x_i| \leq \|\mathbf{x}\| \leq M$. This implies that S is contained in the closed rectangle

$$R = [-M, M] \times \cdots \times [-M, M].$$

Let $V = \mathbb{R}^n \setminus S$. Since S is closed, V is an open set. Then $\widetilde{\mathscr{A}} = \mathscr{A} \cup \{V\}$ is an open covering of $\mathbb{R}^n$, and hence it is an open covering of R. By Theorem 3.26, R is compact. Thus, there exists $\widetilde{\mathscr{B}} \subset \widetilde{\mathscr{A}}$ which is a finite subcover of R. Then $\mathscr{B} = \widetilde{\mathscr{B}} \setminus \{V\}$ is a finite subcollection of $\mathscr{A}$ that covers S. This proves that S is compact.

Example 3.27

We have shown in Example 3.20 that the set

$$B = \{(x, y, z) \mid x^2 + 4y^2 + 9z^2 \leq 36\}$$

is closed and bounded. Hence, it is compact.

Now we can conclude our main theorem from the Bolzano-Weierstrass theorem and the Heine-Borel theorem.

Theorem 3.28

Let S be a subset of $\mathbb{R}^n$. Then the following are equivalent.

(a) S is sequentially compact.

(b) S is closed and bounded.

(c) S is compact.

Remark 3.5

Henceforth, when we say a subset S of $\mathbb{R}^n$ is compact, we mean it is a closed and bounded set, and it is sequentially compact. By Theorem 3.19, a subset S of $\mathbb{R}^n$ has compact closure if and only if it is a bounded set.

Finally, we can conclude the following, which says that continuous functions preserve compactness.

Theorem 3.29

Let $\mathfrak{D}$ be a compact subset of $\mathbb{R}^n$. If the function $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is continuous, then $\mathbf{F}(\mathfrak{D})$ is a compact subset of $\mathbb{R}^m$.

Proof

Since $\mathfrak{D}$ is compact, it is sequentially compact. By Theorem 3.24, $\mathbf{F}(\mathfrak{D})$ is a sequentially compact subset of $\mathbb{R}^m$. Hence, $\mathbf{F}(\mathfrak{D})$ is a compact subset of $\mathbb{R}^m$.

Exercises 3.3

Question 1

Determine whether the following subsets of $\mathbb{R}^2$ is sequentially compact.

(a) $A = \{(x, y) \,|\, x^2 + y^2 = 9\}$.

(b) $B = \{(x, y) \,|\, 0 < x^2 + 4y^2 \leq 36\}$.

(c) $C = \{(x, y) \,|\, x \geq 0, 0 \leq y \leq x^2\}$.

Question 2

Determine whether the following subsets of $\mathbb{R}^3$ is compact.

(a) $A = \{(x, y, z) \,|\, 1 \leq x \leq 2\}$.

(b) $B = \{(x, y, z) \,|\, |x| + |y| + |z| \leq 10\}$.

(c) $C = \{(x, y, z) \,|\, 4 \leq x^2 + y^2 + z^2 \leq 9\}$.

Question 3

Given that A is a compact subset of $\mathbb{R}^n$ and B is a subset of A, show that B is compact if and only if it is closed.

Question 4

If $S_1, \ldots, S_k$ are compact subsets of $\mathbb{R}^n$, show that $S = S_1 \cup \cdots \cup S_n$ is also compact.

Question 5

If A is a compact subset of $\mathbb{R}^m$, B is a compact subset of $\mathbb{R}^n$, show that $A \times B$ is a compact subset of $\mathbb{R}^{m+n}$.

3.4 Applications of Compactness

In this section, we consider the applications of compactness. We are going to use repeatedly the fact that a subset S of $\mathbb{R}^n$ is compact if and only if it is closed and bounded, if and only if it is sequentially compact.

3.4.1 The Extreme Value Theorem

First we define bounded functions.

Definition 3.15 Bounded Functions

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$, and let $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ be a function defined on $\mathfrak{D}$. We say that the function $\mathbf{F}$ is bounded if the set $\mathbf{F}(\mathfrak{D})$ is a bounded subset of $\mathbb{R}^m$. In other words, the function $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is bounded if there is positive number M such that

$$\|\mathbf{F}(\mathbf{x})\| \leq M \qquad \text{for all } \mathbf{x} \in \mathfrak{D}.$$

Example 3.28

Let $\mathfrak{D} = \{(x, y, z) \,|\, 0 < x^2 + y^2 + z^2 < 4\}$, and let $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^2$ be the function defined as

$$\mathbf{F}(x, y, z) = \left(\frac{1}{x^2 + y^2 + z^2}, x + y + z \right).$$

For $k \in \mathbb{Z}^+$, the point $\mathbf{u}_k = (1/k, 0, 0)$ is in $\mathfrak{D}$ and

$$\mathbf{F}(\mathbf{u}_k) = \left(k^2, \frac{1}{k} \right).$$

Thus, $\|\mathbf{F}(\mathbf{u}_k)\| \geq k^2$. This shows that $\mathbf{F}$ is not bounded, even though $\mathfrak{D}$ is a bounded set.

Theorem 3.24 gives the following.

Theorem 3.30

Let $\mathfrak{D}$ be a compact subset of $\mathbb{R}^n$. If the function $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is continuous, then it is bounded.

Proof

By Theorem 3.29, $\mathbf{F}(\mathfrak{D})$ is compact. Hence, it is bounded.

Example 3.29

Let $\mathfrak{D} = \{(x, y, z) \,|\, 1 < x^2 + y^2 + z^2 < 4\}$, and let $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^2$ be the function defined as

$$\mathbf{F}(x, y, z) = \left(\frac{1}{x^2 + y^2 + z^2}, x + y + z \right).$$

Show that $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^2$ is a bounded function.

Solution

Notice that the set $\mathfrak{D}$ is not closed. Therefore, we cannot apply Theorem 3.30 directly. Consider the set $\mathcal{U} = \{(x, y, z) \,|\, 1 \leq x^2 + y^2 + z^2 \leq 4\}$. For any $\mathbf{u} = (x, y, z)$ in $\mathcal{U}$, $\|\mathbf{u}\| \leq 2$. Hence, $\mathcal{U}$ is bounded. The function $f : \mathbb{R}^3 \to \mathbb{R}$ defined as $f(x, y, z) = x^2 + y^2 + z^2$ is continuous, and $\mathcal{U} = f^{-1}([1, 4])$. Since $[1, 4]$ is closed in $\mathbb{R}$, $\mathcal{U}$ is closed in $\mathbb{R}^3$. Since $f(x, y, z) \neq 0$ on $\mathcal{U}$,

$$F_1(x, y, z) = \frac{1}{x^2 + y^2 + z^2}$$

is continuous on $\mathcal{U}$. Being a polynomial function, $F_2(x, y, z) = x + y + z$ is continuous. Thus, $\mathbf{F} : \mathcal{U} \to \mathbb{R}^2$ is continuous. Since $\mathcal{U}$ is closed and bounded, Theorem 3.30 implies that $\mathbf{F} : \mathcal{U} \to \mathbb{R}^2$ is bounded. Since $\mathfrak{D} \subset \mathcal{U}$, $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^2$ is also a bounded function.

Recall that if S is a subset of $\mathbb{R}$, S has maximum value if and only if S is bounded above and $\sup S$ is in S; while S has minimum value if and only if S is bounded below and $\inf S$ is in S.

Definition 3.16 Extremizer and Extreme Values

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$, and let $f : \mathfrak{D} \to \mathbb{R}$ be a function defined on $\mathfrak{D}$.

1. The function f has maximum value if there is a point $\mathbf{x}_0$ in $\mathfrak{D}$ such that

$$f(\mathbf{x}_0) \geq f(\mathbf{x}) \qquad \text{for all } \mathbf{x} \in \mathfrak{D}.$$

   The point $\mathbf{x}_0$ is called a maximizer of f; and $f(\mathbf{x}_0)$ is the maximum value of f.

2. The function f has minimum value if there is a point $\mathbf{x}_0$ in $\mathfrak{D}$ such that

$$f(\mathbf{x}_0) \leq f(\mathbf{x}) \qquad \text{for all } \mathbf{x} \in \mathfrak{D}.$$

   The point $\mathbf{x}_0$ is called a minimizer of f; and $f(\mathbf{x}_0)$ is the minimum value of f.

We have proved in volume I that a sequentially compact subset of $\mathbb{R}$ has a maximum value and a minimum value. This gives us the extreme value theorem.

Theorem 3.31 Extreme Value Theorem

Let $\mathfrak{D}$ be a compact subset of $\mathbb{R}^n$. If the function $f : \mathfrak{D} \to \mathbb{R}$ is continuous, then it has a maximum value and a minimum value.

Proof

By Theorem 3.24, $f(\mathfrak{D})$ is a sequentially compact subset of $\mathbb{R}$. Therefore, f has a maximum value and a minimum value.

Example 3.30

Let $\mathfrak{D} = \{(x, y) \,|\, x^2 + 2x + y^2 \leq 3\}$, and let $f : \mathfrak{D} \to \mathbb{R}$ be the function defined by

$$f(x, y) = x^2 + xy^3 + e^{x-y}.$$

Show that f has a maximum value and a minimum value.

Solution

Notice that

$$\mathfrak{D} = \{(x, y) \,|\, x^2 + 2x + y^2 \leq 3\} = \{(x, y) \,|\, (x + 1)^2 + y^2 \leq 4\}$$

is a closed ball. Thus, it is closed and bounded. The function $f_1(x, y) = x^2 + xy^3$ and the function $g(x, y) = x - y$ are polynomial functions. Hence, they are continuous. The exponential function $h(x) = e^x$ is continuous. Hence, the function $f_2(x, y) = (h \circ g)(x, y) = e^{x-y}$ is continuous. Since $f = f_1 + f_2$, the function $f : \mathfrak{D} \to \mathbb{R}$ is continuous. Since $\mathfrak{D}$ is compact, the function $f : \mathfrak{D} \to \mathbb{R}$ has a maximum value and a minimum value.

Remark 3.6 Extreme Value Property

Let S be a subset of $\mathbb{R}^n$. We say that S has *extreme value property* provided that whenever $f : S \to \mathbb{R}$ is a continuous function, then f has maximum and minimum values.

The extreme value theorem says that if S is compact, then it has extreme value property. Now let us show the converse. Namely, if S has extreme value property, then it is compact, or equivalently, it is closed and bounded. If S is not bounded, the function $f : S \to \mathbb{R}$, $f(\mathbf{x}) = \|\mathbf{x}\|$ is continuous, but it does not have maximum value. If S is not closed, there is a sequence $\{\mathbf{x}_k\}$ in S that converges to a point $\mathbf{x}_0$ that is not in S. The function $g : S \to \mathbb{R}$, $g(\mathbf{x}) = \|\mathbf{x} - \mathbf{x}_0\|$ is continuous and $g(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in S$. Since $\lim_{k\to\infty} g(\mathbf{x}_k) = 0$, we find that $\inf g(S) = 0$. Since $\mathbf{x}_0$ is not in S, there is no point $\mathbf{x}$ in S such that $g(\mathbf{x}) = 0$. Hence, g does not have minimum value. This shows that for S to have extreme value property, it is necessary that S is closed and bounded.

Therefore, a subset S of $\mathbb{R}^n$ has extreme value property if and only if it is compact.

3.4.2 Distance Between Sets

The distance between two sets is defined in the following way.

Definition 3.17 Distance Between Two Sets

Let A and B be two subsets of $\mathbb{R}^n$. The distance between A and B is defined as

$$d(A, B) = \inf \{d(\mathbf{a}, \mathbf{b}) \,|\, \mathbf{a} \in A, \mathbf{b} \in B\}.$$

The distance between two sets is always well-defined and nonnegative. If A and B are not disjoint, then their distance is 0.

Example 3.31

Let $A = \{(x, y) \,|\, x^2 + y^2 < 1\}$ and let $B = [1, 3] \times [-1, 1]$. Find the distance between the two sets A and B.

Solution

For $k \in \mathbb{Z}^+$, let $\mathbf{a}_k$ be the point in A given by

$$\mathbf{a}_k = \left(1 - \frac{1}{k}, 0\right).$$

Let $\mathbf{b} = (1, 0)$. Then $\mathbf{b}$ is in B. Notice that

$$d(\mathbf{a}_k, \mathbf{b}) = \|\mathbf{a}_k - \mathbf{b}\| = \frac{1}{k}.$$

Hence, $d(A, B) \leq \dfrac{1}{k}$ for all $k \in \mathbb{Z}^+$. This shows that the distance between A and B is 0.

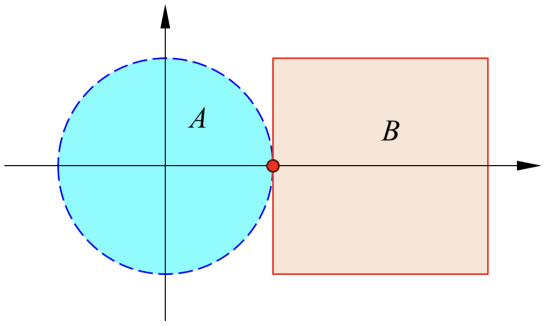


Figure 3.15: The sets A and B in Example 3.31.

In Example 3.31, we find that the distance between two disjoint sets can be 0, even though they are both bounded.

Example 3.32

Let $A = \{(x, y) \,|\, y = 0\}$ and let $B = \{(x, y) \,|\, xy = 1\}$. Find the distance between the two sets A and B.

Solution

For $k \in \mathbb{Z}^+$, let $\mathbf{a}_k = (k, 0)$ and $\mathbf{b}_k = (k, 1/k)$. Then $\mathbf{a}_k$ is in A and $\mathbf{b}_k$ is in B. Notice that

$$d(\mathbf{a}_k, \mathbf{b}_k) = \|\mathbf{a}_k - \mathbf{b}_k\| = \frac{1}{k}.$$

Hence, $d(A, B) \leq \dfrac{1}{k}$ for all $k \in \mathbb{Z}^+$. This shows that the distance between A and B is 0.

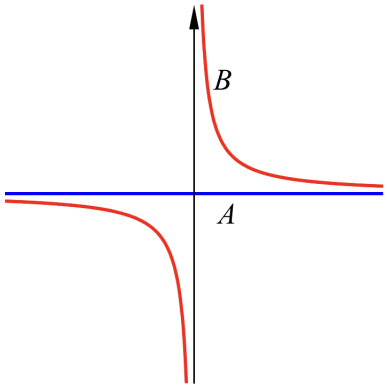


Figure 3.16: The sets A and B in Example 3.32.

In Example 3.32, we find that the distance between two disjoint sets can be 0, even though both of them are closed.

When B is the one-point set $\{\mathbf{x}_0\}$, the distance between A and B is the distance from the point $\mathbf{x}_0$ to the set A. We denote this distance as $d(\mathbf{x}_0, A)$. In other words,

$$d(\mathbf{x}_0, A) = \inf \{d(\mathbf{a}, \mathbf{x}_0) \,|\, \mathbf{a} \in A\}.$$

If $\mathbf{x}_0$ is a point in A, then $d(\mathbf{x}_0, A) = 0$. However, the distance from a point $\mathbf{x}_0$ to a set A can be 0 even though $\mathbf{x}_0$ is not in A. For example, the distance between

the point $\mathbf{x}_0 = (1, 0)$ and the set $A = \{(x, y) \mid x^2 + y^2 < 1\}$ is 0, even thought $\mathbf{x}_0$ is not in A. The following proposition says that this cannot happen if A is closed.

Proposition 3.32

Let A be a closed subset of $\mathbb{R}^n$ and let $\mathbf{x}_0$ be a point in $\mathbb{R}^n$. Then $d(\mathbf{x}_0, A) = 0$ if and only if $\mathbf{x}_0$ is in A.

Proof

If $\mathbf{x}_0$ is in A, it is obvious that $d(\mathbf{x}_0, A) = 0$.
Conversely, if $\mathbf{x}_0$ is not in A, $\mathbf{x}_0$ is in the open set $\mathbb{R}^n \setminus A$. Therefore, there is an $r > 0$ such that $B(\mathbf{x}_0, r) \subset \mathbb{R}^n \setminus A$. For any $\mathbf{a} \in A$, $\mathbf{a} \notin B(\mathbf{x}_0, r)$. Therefore, $\|\mathbf{x}_0 - \mathbf{a}\| \geq r$. Taking infimum over $\mathbf{a} \in A$, we find that $d(\mathbf{x}_0, A) \geq r$. Hence, $d(\mathbf{x}_0, A) \neq 0$.

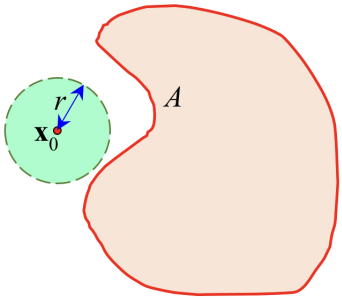


Figure 3.17: A point outside a closed set has positive distance from the set.

Proposition 3.33

Given a subset A of $\mathbb{R}^n$, define the function $f : \mathbb{R}^n \to \mathbb{R}$ by

$$f(\mathbf{x}) = d(\mathbf{x}, A).$$

Then f is a continuous function.

Proof

We prove something stronger. For any $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$, we claim that

$$|f(\mathbf{u}) - f(\mathbf{v})| \leq \|\mathbf{u} - \mathbf{v}\|.$$

This means that f is a Lipschitz function with Lipschitz constant 1, which implies that it is continuous.
Given $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$, if $\mathbf{a}$ is in A, then

$$d(\mathbf{u}, A) \leq \|\mathbf{u} - \mathbf{a}\| \leq \|\mathbf{v} - \mathbf{a}\| + \|\mathbf{u} - \mathbf{v}\|.$$

This shows that

$$\|\mathbf{v} - \mathbf{a}\| \geq d(\mathbf{u}, A) - \|\mathbf{u} - \mathbf{v}\|.$$

Taking infimum over $\mathbf{a} \in A$, we find that

$$d(\mathbf{v}, A) \geq d(\mathbf{u}, A) - \|\mathbf{u} - \mathbf{v}\|.$$

Therefore,

$$f(\mathbf{u}) - f(\mathbf{v}) \leq \|\mathbf{u} - \mathbf{v}\|.$$

Interchanging $\mathbf{u}$ and $\mathbf{v}$, we obtain

$$f(\mathbf{v}) - f(\mathbf{u}) \leq \|\mathbf{u} - \mathbf{v}\|.$$

This proves that

$$|f(\mathbf{u}) - f(\mathbf{v})| \leq \|\mathbf{u} - \mathbf{v}\|.$$

Now we can prove the following.

Theorem 3.34

Let A and C be disjoint subsets of $\mathbb{R}^n$. If A is compact and C is closed, then the distance between A and C is positive.

Proof

By Proposition 3.33, the function $f : A \to \mathbb{R}$, $f(\mathbf{a}) = d(\mathbf{a}, C)$ is continuous. Since A is compact, f has a minimum value. Namely, there is a point $\mathbf{a}_0$ in A such that

$$d(\mathbf{a}_0, C) \leq d(\mathbf{a}, C) \qquad \text{for all } \mathbf{a} \in A.$$

For any $\mathbf{a}$ in A and $\mathbf{c} \in C$,

$$d(\mathbf{a}, \mathbf{c}) \geq d(\mathbf{a}, C) \geq d(\mathbf{a}_0, C).$$

Taking infimum over all $\mathbf{a} \in A$ and $\mathbf{c} \in C$, we find that

$$d(A, C) \geq d(\mathbf{a}_0, C).$$

By definition, we also have $d(A, C) \leq d(\mathbf{a}_0, C)$. Thus, $d(A, C) = d(\mathbf{a}_0, C)$. Since A and C are disjoint and C is closed, Proposition 3.32 implies that $d(A, C) = d(\mathbf{a}_0, C) > 0$.

An equivalent form of Theorem 3.34 is the following important theorem.

Theorem 3.35

Let A be a compact subset of $\mathbb{R}^n$, and let U be an open subset of $\mathbb{R}^n$ that contains A. Then there is a positive number δ such that if $\mathbf{x}$ is a point in $\mathbb{R}^n$ that has a distance less than δ from the set A, then $\mathbf{x}$ is in U.

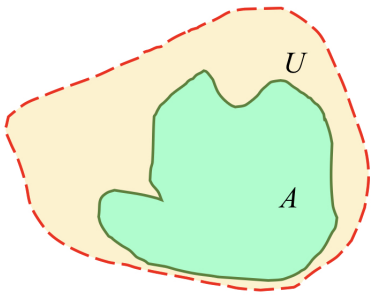


Figure 3.18: A compact set has a positive distance from the boundary of the open set that contains it.

Proof

Let $C = \mathbb{R}^n \setminus U$. Then C is a closed subset of $\mathbb{R}^n$ that is disjoint from A. By Theorem 3.34, $\delta = d(A, C) > 0$. If $\mathbf{x}$ is in $\mathbb{R}^n$ and $d(\mathbf{x}, A) < \delta$, then $\mathbf{x}$ cannot be in C. Therefore, $\mathbf{x}$ is in U.

As a corollary, we have the following.

Corollary 3.36

Let A be a compact subset of $\mathbb{R}^n$, and let U be an open subset of $\mathbb{R}^n$ that contains A. Then there is a positive number r and a compact set K such that $A \subset K \subset U$, and if $\mathbf{x}$ is a point in $\mathbb{R}^n$ that has a distance less than r from the set A, then $\mathbf{x}$ is in K.

Proof

By Theorem 3.35, there is a positive number δ such that if $\mathbf{x}$ is a point in $\mathbb{R}^n$ that has a distance less than δ from the set A, then $\mathbf{x}$ is in U. Take $r = \delta/2$, and let

$$K = \overline{V}, \qquad \text{where } V = \bigcup_{\mathbf{u} \in A} B(\mathbf{u}, r).$$

Since A is compact, it is bounded. There is a positive number M such that $\|\mathbf{u}\| \leq M$ for all $\mathbf{u} \in A$. If $\mathbf{x} \in V$, then there is an $\mathbf{u} \in A$ such that $\|\mathbf{x} - \mathbf{u}\| < r$. This implies that $\|\mathbf{x}\| \leq M + r$. Hence, the set V is also bounded. Since K is the closure of a bounded set, K is compact. Since $A \subset V$, $A \subset K$. If $\mathbf{w} \in K$, since K is the closure of V, there is a point $\mathbf{v}$ in V that lies in $B(\mathbf{w}, r)$. By the definition of V, there is a point $\mathbf{u}$ in A such that $\mathbf{v} \in B(\mathbf{u}, r)$. Thus,

$$\|\mathbf{w} - \mathbf{u}\| \leq \|\mathbf{w} - \mathbf{v}\| + \|\mathbf{v} - \mathbf{u}\| < r + r = \delta.$$

This implies that $\mathbf{w}$ has a distance less than δ from A. Hence, $\mathbf{w}$ is in U. This shows that $K \subset U$.

Now if $\mathbf{x}$ is a point that has distance d less than r from the set A, there is a point $\mathbf{u}$ is A such that $\|\mathbf{x} - \mathbf{u}\| < r$. This implies that $\mathbf{x} \in B(\mathbf{u}, r) \in V \subset K$.

3.4.3 Uniform Continuity

In Section 2.4, we have discussed uniform continuity. Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$ and let $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ be a function defined on $\mathfrak{D}$. We say that $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is uniformly continuous provided that for any $\varepsilon > 0$, there exists $\delta > 0$ such that for any points $\mathbf{u}$ and $\mathbf{v}$ in $\mathfrak{D}$, if $\|\mathbf{u} - \mathbf{v}\| < \delta$, then $\|\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v})\| < \varepsilon$. If a function is uniformly continuous, it is continuous. The converse is not true. However, a continuous function that is defined on a compact subset of $\mathbb{R}^n$ is uniformly continuous. This is an important theorem in analysis.

Theorem 3.37

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$, and let $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ be a continuous function defined on $\mathfrak{D}$. If $\mathfrak{D}$ is compact, then $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is uniformly continuous.

Proof

Assume to the contrary that $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is not uniformly continuous. Then there exists an $\varepsilon > 0$, for any $\delta > 0$, there exist points $\mathbf{u}$ and $\mathbf{v}$ in $\mathfrak{D}$ such that $\|\mathbf{u} - \mathbf{v}\| < \delta$ and $\|\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v})\| \geq \varepsilon$. This implies that for any $k \in \mathbb{Z}^+$, there exist $\mathbf{u}_k$ and $\mathbf{v}_k$ in $\mathfrak{D}$ such that $\|\mathbf{u}_k - \mathbf{v}_k\| < 1/k$ and $\|\mathbf{F}(\mathbf{u}_k) - \mathbf{F}(\mathbf{v}_k)\| \geq \varepsilon$. Since $\mathfrak{D}$ is sequentially compact, there is a subsequence $\{\mathbf{u}_{k_j}\}$ of $\{\mathbf{u}_k\}$ that converges to a point $\mathbf{u}_0$ in $\mathfrak{D}$. Consider the sequence $\{\mathbf{v}_{k_j}\}$ in $\mathfrak{D}$. It has a subsequence $\{\mathbf{v}_{k_{j_l}}\}$ that converges to a point $\mathbf{v}_0$ in $\mathfrak{D}$. Being a subsequence of $\{\mathbf{u}_{k_j}\}$, the sequence $\{\mathbf{u}_{k_{j_l}}\}$ also converges to $\mathbf{u}_0$.

Since $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is continuous, the sequences $\{\mathbf{F}(\mathbf{u}_{k_{j_l}})\}$ and $\{\mathbf{F}(\mathbf{v}_{k_{j_l}})\}$ converge to $\mathbf{F}(\mathbf{u}_0)$ and $\mathbf{F}(\mathbf{v}_0)$ respectively. Notice that by construction,

$$\|\mathbf{F}(\mathbf{u}_{k_{j_l}}) - \mathbf{F}(\mathbf{v}_{k_{j_l}})\| \geq \varepsilon \qquad \text{for all } l \in \mathbb{Z}^+.$$

Thus, $\|\mathbf{F}(\mathbf{u}_0) - \mathbf{F}(\mathbf{v}_0)\| \geq \varepsilon$. This implies that $\mathbf{F}(\mathbf{u}_0) \neq \mathbf{F}(\mathbf{v}_0)$, and so $\mathbf{u}_0 \neq \mathbf{v}_0$.

Since $k_{j_1}, k_{j_2}, \ldots$ is a strictly increasing sequence of positive integers, $k_{j_l} \geq l$. Thus,

$$\|\mathbf{u}_{k_{j_l}} - \mathbf{v}_{k_{j_l}}\| < \frac{1}{k_{j_l}} \leq \frac{1}{l}.$$

Taking $l \to \infty$ implies that $\mathbf{u}_0 = \mathbf{v_0}$. This gives a contradiction. Thus, $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ must be uniformly continuous.

Example 3.33

Let $\mathfrak{D} = (-1, 4) \times (-7, 5]$ and let $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^3$ be the function defined as

$$\mathbf{F}(x, y) = \left(\sin(x + y), \sqrt{x + y + 8}, e^{xy}\right).$$

Show that $\mathbf{F}$ is uniformly continuous.

Solution

Let $\mathcal{U} = [-1, 4] \times [-7, 5]$. Then $\mathcal{U}$ is a closed and bounded subset of $\mathbb{R}^2$ that contains $\mathfrak{D}$. The functions $f_1(x, y) = x + y$, $f_2(x, y) = x + y + 8$ and $f_3(x, y) = xy$ are polynomial functions. Hence, they are continuous. If $(x, y) \in \mathcal{U}$, $x \geq -1$, $y \geq -7$ and so $f_2(x, y) = x + y + 8 \geq 0$. Thus, $f_2(\mathcal{U})$ is contained in the domain of the square root function. Since the square root function, the sine function and the exponential function are continuous on their domains, we find that the functions

$$F_1(x, y) = \sin(x + y), \quad F_2(x, y) = \sqrt{x + y + 8}, \quad F_3(x, y) = e^{xy}$$

are continuous on $\mathcal{U}$. Since $\mathcal{U}$ is closed and bounded, $\mathbf{F} : \mathcal{U} \to \mathbb{R}^3$ is uniformly continuous. Since $\mathfrak{D} \subset \mathcal{U}$, $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^3$ is uniformly continuous.

3.4.4 Linear Transformations and Quadratic Forms

In Chapter 2, we have seen that a linear transformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^m$ is a matrix transformation. Namely, there exists an $m \times n$ matrix such that

$$\mathbf{T}(\mathbf{x}) = A\mathbf{x} \qquad \text{for all } \mathbf{x} \in \mathbb{R}^n.$$

A linear transformation is continuous. Theorem 2.34 says that a linear transformation is Lipschitz. More precisely, there exists a positive constant $c > 0$ such that

$$\|\mathbf{T}(\mathbf{x})\| \leq c\|\mathbf{x}\| \qquad \text{for all } \mathbf{x} \in \mathbb{R}^n.$$

Theorem 2.5 says that when $m = n$, a linear transformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$ is invertible if and only if it is one-to-one, if and only if the matrix A is invertible, if and only if $\det A \neq 0$. Here we want to give a stronger characterization of a linear transformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$ that is invertible.

Recall that to show that a linear transformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^m$ is one-to-one, it is sufficient to show that $\mathbf{T}(\mathbf{x}) = \mathbf{0}$ implies that $\mathbf{x} = \mathbf{0}$.

Theorem 3.38

Let $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$ be a linear transformation. The following are equivalent.

(a) $\mathbf{T}$ is invertible.

(b) There is a positive constant a such that

$$\|\mathbf{T}(\mathbf{x})\| \geq a\|\mathbf{x}\| \qquad \text{for all } \mathbf{x} \in \mathbb{R}^n.$$

Proof

(b) implies (a) is easy. Notice that (b) says that

$$\|\mathbf{x}\| \leq \frac{1}{a}\|\mathbf{T}(\mathbf{x})\| \qquad \text{for all } \mathbf{x} \in \mathbb{R}^n. \tag{3.3}$$

If $\mathbf{T}(\mathbf{x}) = \mathbf{0}$, then $\|\mathbf{T}(\mathbf{x})\| = 0$. Eq. (3.3) implies that $\|\mathbf{x}\| = 0$. Thus, $\mathbf{x} = \mathbf{0}$. This proves that $\mathbf{T}$ is one-to-one. Hence, it is invertible.
Conversely, assume that $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$ is invertible. Let

$$S^{n-1} = \left\{(x_1, \ldots, x_n) \,|\, x_1^2 + \cdots + x_n^2 = 1\right\}$$

be the standard unit $(n-1)$-sphere in $\mathbb{R}^n$. We have seen that S^{n-1} is compact. For any $\mathbf{u} \in S^{n-1}$, $\mathbf{u} \neq \mathbf{0}$. Therefore, $\mathbf{T}(\mathbf{u}) \neq \mathbf{0}$ and so $\|\mathbf{T}(\mathbf{u})\| > 0$. The function $f : S^{n-1} \to \mathbb{R}^n$, $f(\mathbf{u}) = \|\mathbf{T}(\mathbf{u})\|$ is continuous. Hence, it has a mimimum value at some $\mathbf{u}_0$ on S^{n-1}. Let $a = \|\mathbf{T}(\mathbf{u}_0)\|$. Then $a > 0$. Since a is the minimum value of f,

$$\|\mathbf{T}(\mathbf{u})\| \geq a \qquad \text{for all } \mathbf{u} \in S^{n-1}.$$

Notice that if $\mathbf{x} = \mathbf{0}$, $\|\mathbf{T}(\mathbf{x})\| \geq a\|\mathbf{x}\|$ holds trivially. If $\mathbf{x}$ is in $\mathbb{R}^n$ and $\mathbf{x} \neq 0$, let $\mathbf{u} = \alpha\mathbf{x}$, where $\alpha = 1/\|\mathbf{x}\|$. Then $\mathbf{u}$ is in S^{n-1}. Therefore, $\|\mathbf{T}(\mathbf{u})\| \geq a$. Since $\mathbf{T}(\mathbf{u}) = \alpha\mathbf{T}(\mathbf{x})$, and $\alpha > 0$, we find that $\|\mathbf{T}(\mathbf{u})\| = \alpha\|\mathbf{T}(\mathbf{x})\|$. Hence, $\alpha\|\mathbf{T}(\mathbf{x})\| \geq a$. This gives

$$\|\mathbf{T}(\mathbf{x})\| \geq \frac{a}{\alpha} = a\|\mathbf{x}\|.$$

In Section 2.1.5, we have reviewed some theories of quadratic forms from linear algebra. In Theorem 2.7, we state for a quadratic form $Q_A : \mathbb{R}^n \to \mathbb{R}$, $Q_A(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ defined by the symmetric matrix A, we have

$$\lambda_n\|\mathbf{x}\|^2 \leq Q_A(\mathbf{x}) \leq \lambda_1\|\mathbf{x}\|^2 \qquad \text{for all } \mathbf{x} \in \mathbb{R}^n.$$

Here λ_n is the smallest eigenvalue of A, and λ_1 is the largest eigenvalue of A.

We have used Theorem 2.7 to prove that a linear transformation is Lipschitz in Theorem 2.34. It boils down to the fact that if $\mathbf{T}(\mathbf{x}) = A\mathbf{x}$, then $\|\mathbf{T}(\mathbf{x})\|^2 = \mathbf{x}^T(A^TA)\mathbf{x}$, and A^TA is a positive semi-definite quadractic form. In fact, we can also use Theorem 2.7 to prove Theorem 3.38, using the fact that if A is invertible, then A^TA is positive definite.

Let us prove a weaker version of Theorem 2.7 here, which is sufficient to establish Theorem 3.38 and the theorem which says that a linear transformation is Lipschitz.

Theorem 3.39

Let A be an $n \times n$ symmetric matrix, and let $Q_A : \mathbb{R}^n \to \mathbb{R}$ be the quadratic form $Q_A(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ defined by A. There exists constants a and b such that

$$a\|\mathbf{x}\|^2 \leq Q_A(\mathbf{x}) \leq b\|\mathbf{x}\|^2 \qquad \text{for all } \mathbf{x} \in \mathbb{R}^n,$$

$Q_A(\mathbf{u}) = a\|\mathbf{u}\|^2$ and $Q_A(\mathbf{v}) = b\|\mathbf{v}\|^2$ for some $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$. Therefore,

(i) if A is positive semi-definite, $b \geq a \geq 0$;

(ii) if A is positive definite, $b \geq a > 0$.

Proof

As in the proof of Theorem 3.38, consider the continuous function $Q_A : S^{n-1} \to \mathbb{R}$. Since S^{n-1} is compact, there exsits $\mathbf{u}$ and $\mathbf{v}$ in S^{n-1} such that

$$Q_A(\mathbf{u}) \leq Q_A(\mathbf{w}) \leq Q_A(\mathbf{v}) \qquad \text{for all } \mathbf{w} \in S^{n-1}.$$

Let $a = Q_A(\mathbf{u})$ and $b = Q_A(\mathbf{v})$. If $\mathbf{x} = \mathbf{0}$, $a\|\mathbf{x}\|^2 \leq Q_A(\mathbf{x}) \leq b\|\mathbf{x}\|^2$ holds trivially. Now if $\mathbf{x}$ is in $\mathbb{R}^n$ and $\mathbf{x} \neq \mathbf{0}$, let $\mathbf{w} = \alpha\mathbf{x}$, where $\alpha = 1/\|\mathbf{x}\|$. Then $\mathbf{w}$ in on S^{n-1}. Notice that

$$Q_A(\mathbf{w}) = \alpha^2 Q_A(\mathbf{x}).$$

Hence,

$$a \leq \frac{1}{\|\mathbf{x}\|^2} Q_A(\mathbf{x}) \leq b.$$

This proves that

$$a\|\mathbf{x}\|^2 \leq Q_A(\mathbf{x}) \leq b\|\mathbf{x}\|^2.$$

3.4.5 Lebesgue Number Lemma

Now let us prove the following important theorem.

Theorem 3.40 Lebesgue Number Lemma

Let A be a subset of $\mathbb{R}^n$, and let $\mathscr{A} = \{U_\alpha \,|\, \alpha \in J\}$ be an open covering of A. If A is compact, there exists a positive number δ such that if S is a subset of A and diam $S < \delta$, then S is contained in one of the elements of $\mathscr{A}$. Such a positive number δ is called the Lebesgue number of the covering $\mathscr{A}$.

We give two proofs of this theorem.

First Proof of the Lebesgue Number Lemma

We use proof by contradiction. Assume that there does not exist a positive number δ such that any subset S of A that has diameter less than δ lies inside an open set in $\mathscr{A}$. Then for any $k \in \mathbb{Z}^+$, there is a subset S_k of A whose diameter is less than $1/k$, but S_k is not contained in any element of $\mathscr{A}$.

For each $k \in \mathbb{Z}^+$, the set S_k cannot be empty. Let $\mathbf{x}_k$ be any point in S_k. Then $\{\mathbf{x}_k\}$ is a sequence of points in A. Since A is sequentially compact, there is a subsequence $\{\mathbf{x}_{k_m}\}$ that converges to a point $\mathbf{x}_0$ in A. Since $\mathscr{A}$ is an open covering of A, there exists $\beta \in J$ such that $\mathbf{x}_0 \in U_\beta$. Since U_β is open, there exists $r > 0$ such that $B(\mathbf{x}_0, r) \subset U_\beta$. Since the sequence $\{\mathbf{x}_{k_m}\}$ converges $\mathbf{x}_0$, there is a positive integer M such that for all $m \geq M$, $\mathbf{x}_{k_m} \in B(\mathbf{x}_0, r/2)$. There exists an integer $j \geq M$ such that $1/k_j < r/2$. If $\mathbf{x} \in A_{k_j}$, then

$$\|\mathbf{x} - \mathbf{x}_{k_j}\| \leq \operatorname{diam} A_{k_j} < \frac{1}{k_j} < \frac{r}{2}.$$

Since $\mathbf{x}_{k_j} \in B(\mathbf{x}_0, r/2)$, $\|\mathbf{x}_{k_j} - \mathbf{x}_0\| < r/2$. Therefore, $\|\mathbf{x} - \mathbf{x}_0\| < r$. This proves that $\mathbf{x} \in B(\mathbf{x}_0, r) \subset U_\beta$. Thus, we have shown that $A_{k_j} \subset U_\beta$. But this contradicts to A_{k_j} does not lie in any element of $\mathscr{A}$.

Second Proof of the Lebesgue Number Lemma

Since A is compact, there are finitely many indices $\alpha_1, \ldots, \alpha_m$ in J such that

$$A \subset \bigcup_{j=1}^{m} U_{\alpha_j}.$$

For $1 \leq j \leq m$, let $C_j = \mathbb{R}^n \setminus U_{\alpha_j}$. Then C_j is a closed set and $\bigcap_{j=1}^{m} C_j$ is disjoint from A. By Theorem 3.33, the function $f_j : A \to \mathbb{R}$, $f_j(\mathbf{x}) = d(\mathbf{x}, C_j)$ is continuous. Define $f : A \to \mathbb{R}$ by

$$f(\mathbf{x}) = \frac{f_1(\mathbf{x}) + \cdots + f_m(\mathbf{x})}{m}.$$

Then f is also a continuous function. Since A is compact, there is a point $\mathbf{a}_0$ in A such that

$$f(\mathbf{a}_0) \leq f(\mathbf{a}) \qquad \text{for all } \mathbf{a} \in A.$$

Notice that $f_j(\mathbf{a}_0) \geq 0$ for all $1 \leq j \leq m$. Since $\bigcap_{j=1}^{m} C_j$ is disjoint from A, there is an $1 \leq k \leq m$ such that $\mathbf{a}_0 \notin C_k$. Proposition 3.32 says that $f_k(\mathbf{a}_0) = d(\mathbf{a}_0, C_k) > 0$. Hence, $f(\mathbf{a}_0) > 0$. Let $\delta = f(\mathbf{a}_0)$. It is the minimum value of the function $f : A \to \mathbb{R}$.
Now let S be a nonempty subset of A such that $\operatorname{diam} S < \delta$. Take a point $\mathbf{x}_0$ in S. Let $1 \leq l \leq m$ be an integer such that

$$f_l(\mathbf{x}_0) \geq f_j(\mathbf{x}_0) \qquad \text{for all } 1 \leq j \leq m.$$

Then

$$\delta \leq f(\mathbf{x}_0) \leq f_l(\mathbf{x}_0) = d(\mathbf{x}_0, C_l).$$

For any $\mathbf{u} \in C_l$,

$$d(\mathbf{x}_0, \mathbf{u}) \geq d(\mathbf{x}_0, C_l) \geq \delta.$$

If $\mathbf{x} \in S$, then $d(\mathbf{x}, \mathbf{x}_0) \leq \operatorname{diam} S < \delta$. This implies that $\mathbf{x}$ is not in C_l. Hence, it must be in U_{α_l}. This shows that S is contained in U_{α_l}, which is an element of $\mathscr{A}$. This completes the proof of the theorem.

The Lebesgue number lemma can be used to give an alternative proof of Theorem 3.37, which says that a continuous function defined on a compact subset of $\mathbb{R}^n$ is uniformly continuous.

Alternative Proof of Theorem 3.37

Fixed $\varepsilon > 0$. We want to show that there exists $\delta > 0$ such that if $\mathbf{u}$ and $\mathbf{v}$ are in $\mathfrak{D}$ and $\|\mathbf{u} - \mathbf{v}\| < \delta$, then $\|\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v})\| < \varepsilon$.

We will construct an open covering of $\mathfrak{D}$ indexed by $J = \mathfrak{D}$. Since $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is continuous, for each $\mathbf{x} \in \mathfrak{D}$, there is a positive number $\delta_{\mathbf{x}}$ (depending on $\mathbf{x}$), such that if $\mathbf{u}$ is in $\mathfrak{D}$ and $\|\mathbf{u} - \mathbf{x}\| < \delta_{\mathbf{x}}$, then $\|\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{x})\| < \varepsilon/2$. Let $U_{\mathbf{x}} = B(\mathbf{x}, \delta_{\mathbf{x}})$. Then $U_{\mathbf{x}}$ is an open set. If $\mathbf{u}$ and $\mathbf{v}$ are in $U_{\mathbf{x}}$, $\|\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{x})\| < \varepsilon/2$ and $\|\mathbf{F}(\mathbf{v}) - \mathbf{F}(\mathbf{x})\| < \varepsilon/2$. Thus, $\|\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v})\| < \varepsilon$.

Now $\mathscr{A} = \{U_{\mathbf{x}} \,|\, \mathbf{x} \in \mathfrak{D}\}$ is an open covering of $\mathfrak{D}$. Since $\mathfrak{D}$ is compact, the Lebesgue number lemma implies that there exists a number $\delta > 0$ such that if S is a subset of $\mathfrak{D}$ that has diameter less than δ, then S is contained in one of the $U_{\mathbf{x}}$ for some $\mathbf{x} \in \mathfrak{D}$. We claim that this is the δ that we need.

If $\mathbf{u}$ and $\mathbf{v}$ are two points in $\mathfrak{D}$ and $\|\mathbf{u} - \mathbf{v}\| < \delta$, then $S = \{\mathbf{u}, \mathbf{v}\}$ is a set with diameter less than δ. Hence, there is an $\mathbf{x} \in \mathfrak{D}$ such that $S \subset U_{\mathbf{x}}$. This implies that $\mathbf{u}$ and $\mathbf{v}$ are in $U_{\mathbf{x}}$. Hence, $\|\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v})\| < \varepsilon$. This completes the proof.

Exercises 3.4

Question 1

Let $\mathfrak{D} = \{(x, y) \,|\, 2 < x^2 + 4y^2 < 10\}$, and let $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^3$ be the function defined as

$$\mathbf{F}(x, y) = \left(\frac{x}{x^2 + y^2}, \frac{y}{x^2 + y^2}, \frac{x^2 - y^2}{x^2 + y^2} \right).$$

Show that the function $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^3$ is bounded.

Question 2

Let $\mathfrak{D} = \{(x, y, z) \,|\, 1 \leq x^2 + 4y^2 \leq 10, 0 \leq z \leq 5\}$, and let $f : \mathfrak{D} \to \mathbb{R}$ be the function defined as

$$f(x, y, z) = \frac{x^2 - y^2}{x^2 + y^2 + z^2}.$$

Show that the function $f : \mathfrak{D} \to \mathbb{R}$ has a maximum value and a minimum value.

Question 3

Let $A = \{(x, y) \,|\, x^2 + 4y^2 \leq 16\}$ and $B = \{(x, y) \,|\, x + y \geq 10\}$. Show that the distance between the sets A and B is positive.

Question 4

Let $\mathfrak{D} = \{(x, y, z) \,|\, x^2 + y^2 + z^2 \leq 20\}$ and let $f : \mathfrak{D} \to \mathbb{R}$ be the function defined as

$$f(x, y, z) = e^{x^2 + 4z^2}.$$

Show that $f : \mathfrak{D} \to \mathbb{R}$ is uniformly continuous.

Question 5

Let $\mathfrak{D} = (-1, 2) \times (-6, 0)$ and let $f : \mathfrak{D} \to \mathbb{R}$ be the function defined as

$$f(x, y) = \sqrt{x + y + 7} + \ln(x^2 + y^2 + 1).$$

Show that $f : \mathfrak{D} \to \mathbb{R}$ is uniformly continuous.

Chapter 4

Differentiating Functions of Several Variables

In this chapter, we study differential calculus of functions of several variables.

4.1 Partial Derivatives

When $f : (a, b) \to \mathbb{R}$ is a function defined on an open interval (a, b), the derivative of the function at a point x_0 in (a, b) is defined as

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h},$$

provided that the limit exists. The derivative gives the instantaneous rate of change of the function at the point x_0. Geometrically, it is the slope of the tangent line to the graph of the function $f : (a, b) \to \mathbb{R}$ at the point $(x_0, f(x_0))$.

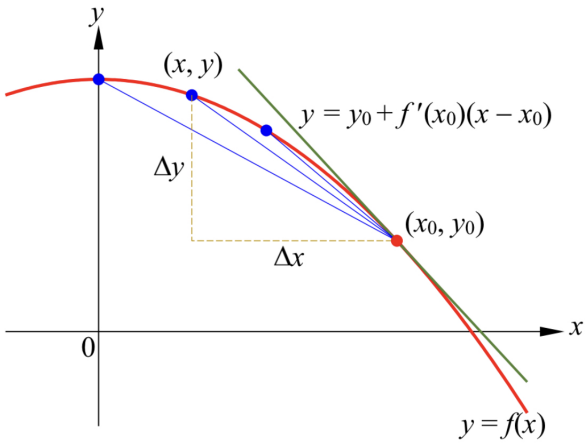


Figure 4.1: Derivative as slope of tangent line.

Now consider a function $f : \mathcal{O} \to \mathbb{R}$ that is defined on an open subset $\mathcal{O}$ of $\mathbb{R}^n$, where $n \geq 2$. What is the natural way to extend the concept of derivatives to this function?

From the perspective of rate of change, we need to consider the change of f in various *different directions*. This leads us to consider directional derivatives. Another perspective is to regard existence of derivatives as *differentiability* and *first-order approximation*. Later we will see that all these are closely related.

First let us consider the rates of change of the function $f : \mathcal{O} \to \mathbb{R}$ at a point $\mathbf{x}_0$ in $\mathcal{O}$ along the directions of the coordinate axes. These are called partial derivatives.

Definition 4.1 Partial Derivatives

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $f : \mathcal{O} \to \mathbb{R}$ be a function defined on $\mathcal{O}$. For $1 \leq i \leq n$, we say that the function $f : \mathcal{O} \to \mathbb{R}$ has a partial derivative with respect to its i^{th} component at the point $\mathbf{x}_0$ if the limit

$$\lim_{h \to 0} \frac{f(\mathbf{x}_0 + h\mathbf{e}_i) - f(\mathbf{x}_0)}{h}$$

exists. In this case, we denote the limit by $\dfrac{\partial f}{\partial x_i}(\mathbf{x}_0)$, and call it the partial derivative of $f : \mathcal{O} \to \mathbb{R}$ with respect to x_i at $\mathbf{x}_0$.

We say that the function $f : \mathcal{O} \to \mathbb{R}$ has partial derivatives at $\mathbf{x}_0$ if $\dfrac{\partial f}{\partial x_i}(\mathbf{x}_0)$ exists for all $1 \leq i \leq n$.

Remark 4.1

When we consider partial derivatives of a function, we always assume that the domain of the function is an open set $\mathcal{O}$, so that each point $\mathbf{x}_0$ in the domain is an interior point of $\mathcal{O}$, and a limit point of $\mathcal{O} \setminus \{\mathbf{x}_0\}$. By definition of open sets, there exists $r > 0$ such that $B(\mathbf{x}_0, r)$ is contained in $\mathcal{O}$. This allows us to compare the function values of f in a neighbourhood of $\mathbf{x}_0$ from various different directions.

By definition, $\dfrac{\partial f}{\partial x_i}(\mathbf{x}_0)$ measures the rate of change of f at $\mathbf{x}_0$ in the direction of $\mathbf{e}_i$. It can also be interpreted as the slope of a curve at the point $(\mathbf{x}_0, f(\mathbf{x}_0))$ on the surface $x_{n+1} = f(\mathbf{x})$, as shown in Figure 4.2

Notations for Partial Derivatives

An alternative notation for $\dfrac{\partial f}{\partial x_i}(\mathbf{x}_0)$ is $f_{x_i}(\mathbf{x}_0)$.

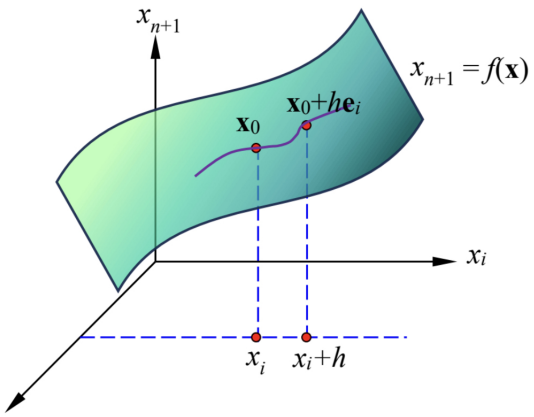


Figure 4.2: Partial derivative.

Remark 4.2 Partial Derivatives

Let $\mathbf{x}_0 = (a_1, a_2, \ldots, a_n)$ and define the function $g : (-r, r) \to \mathbb{R}$ by

$$g(h) = f(\mathbf{x}_0 + h\mathbf{e}_i) = f(a_1, \ldots, a_{i-1}, a_i + h, a_{i+1}, \ldots, a_n).$$

Then

$$\lim_{h\to 0} \frac{f(\mathbf{x}_0 + h\mathbf{e}_i) - f(\mathbf{x}_0)}{h} = \lim_{h\to 0} \frac{g(h) - g(0)}{h} = g'(0).$$

Thus, $f_{x_i}(\mathbf{x}_0)$ exists if and only if $g(h)$ is differentiable at $h = 0$. Moreover, to find $f_{x_i}(\mathbf{x}_0)$, we regard the variables $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ as constants, and differentiate with respect to x_i. Hence, the derivative rules such as sum rule, product rule and quotient rule still work for partial derivatives, as long as one is clear which variable to take derivative, which variable to be regarded as constant.

Example 4.1

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function defined as $f(x, y) = x^2 y$. Find $f_x(1, 2)$ and $f_y(1, 2)$.

Solution

$$\frac{\partial f}{\partial x} = 2xy, \qquad \frac{\partial f}{\partial y} = x^2.$$

Therefore,

$$f_x(1, 2) = 4, \qquad f_y(1, 2) = 1.$$

Example 4.2

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function defined as $f(x, y) = |x + y|$. Determine whether $f_x(0, 0)$ exists.

Solution

By definition, $f_x(0, 0)$ is given by the limit

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

if it exists. Since

$$\lim_{h \to 0} \frac{f(h, 0) - f(0, 0)}{h} = \lim_{h \to 0} \frac{|h|}{h},$$

and

$$\lim_{h \to 0^-} \frac{|h|}{h} = -1 \quad \text{and} \quad \lim_{h \to 0^+} \frac{|h|}{h} = 1,$$

the limit

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

does not exist. Hence, $f_x(0, 0)$ does not exist.

Definition 4.2

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $f : \mathcal{O} \to \mathbb{R}$ be a function defined on $\mathcal{O}$. If the function $f : \mathcal{O} \to \mathbb{R}$ has partial derivative with respect to x_i at every point of $\mathcal{O}$, this defines the function $f_{x_i} : \mathcal{O} \to \mathbb{R}$. In this case, we say that the partial derivative of f with respect to x_i exists.
If $f_{x_i} : \mathcal{O} \to \mathbb{R}$ exists for all $1 \leq i \leq n$, we say that the function $f : \mathcal{O} \to \mathbb{R}$ has partial derivatives.

Example 4.3

Find the partial derivatives of the function $f : \mathbb{R}^3 \to \mathbb{R}$ defined as

$$f(x, y, z) = \sin(xy + z) + \frac{3x}{y^2 + z^2 + 1}.$$

Solution

$$\frac{\partial f}{\partial x}(x, y, z) = y \cos(xy + z) + \frac{3}{y^2 + z^2 + 1},$$
$$\frac{\partial f}{\partial y}(x, y, z) = x \cos(xy + z) - \frac{6xy}{(y^2 + z^2 + 1)^2},$$
$$\frac{\partial f}{\partial z}(x, y, z) = \cos(xy + z) - \frac{6xz}{(y^2 + z^2 + 1)^2}.$$

For a function defined on an open subset of $\mathbb{R}^n$, there are n partial derivatives with respect to the n directions defined by the coordinate axes. These define a vector in $\mathbb{R}^n$.

Definition 4.3 Gradient

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $\mathbf{x}_0$ be a point in $\mathcal{O}$. If the function $f : \mathcal{O} \to \mathbb{R}$ has partial derivatives at $\mathbf{x}_0$, we define the gradient of the function f at $\mathbf{x}_0$ as the vector in $\mathbb{R}^n$ given by

$$\nabla f(\mathbf{x}_0) = \left(\frac{\partial f}{\partial x_1}(\mathbf{x}_0), \frac{\partial f}{\partial x_2}(\mathbf{x}_0), \cdots, \frac{\partial f}{\partial x_n}(\mathbf{x}_0) \right).$$

Let us revisit Example 4.3.

Example 4.4

The gradient of the function $f : \mathbb{R}^3 \to \mathbb{R}$ defined as

$$f(x, y, z) = \sin(xy + z) + \frac{3x}{y^2 + z^2 + 1}$$

in Example 4.3 is the function $\nabla f : \mathbb{R}^3 \to \mathbb{R}^3$,

$$\nabla f(x, y, z) = \begin{bmatrix} y\cos(xy + z) + \dfrac{3}{y^2 + z^2 + 1} \\ x\cos(xy + z) - \dfrac{6xy}{(y^2 + z^2 + 1)^2} \\ \cos(xy + z) - \dfrac{6xz}{(y^2 + z^2 + 1)^2} \end{bmatrix}.$$

In particular,

$$\nabla f(1, -1, 1) = \left(0, \frac{5}{3}, \frac{1}{3}\right).$$

It is straightforward to extend the definition of partial derivative to a function $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ whose codomain is $\mathbb{R}^m$ with $m \geq 2$.

Definition 4.4

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ be a function defined on $\mathcal{O}$. Given $\mathbf{x}_0$ in $\mathcal{O}$ and $1 \leq i \leq n$, we say that $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ has partial derivative with respect to x_i at the point $\mathbf{x}_0$ if the limit

$$\frac{\partial \mathbf{F}}{\partial x_i}(\mathbf{x}_0) = \lim_{h \to 0} \frac{\mathbf{F}(\mathbf{x}_0 + h\mathbf{e}_i) - \mathbf{F}(\mathbf{x}_0)}{h}$$

exists. We say that $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ has partial derivative at the point $\mathbf{x}_0$ if $\dfrac{\partial \mathbf{F}}{\partial x_i}(\mathbf{x}_0)$ exists for each $1 \leq i \leq n$. We say that $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ has partial derivative if it has partial derivative at each point of $\mathcal{O}$.

Since the limit of a function $\mathbf{G} : (-r, r) \to \mathbb{R}^m$ when $h \to 0$ exists if and only if the limit of each component function $G_j : (-r, r) \to \mathbb{R}$, $1 \leq j \leq m$ when $h \to 0$ exists, we have the following.

Proposition 4.1

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ be a function defined on $\mathcal{O}$. Given $\mathbf{x}_0$ in $\mathcal{O}$ and $1 \leq i \leq n$, $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ has partial derivative with respect to x_i at the point $\mathbf{x}_0$ if and only if if each component function $F_j : \mathcal{O} \to \mathbb{R}$, $1 \leq j \leq m$ has partial derivative with respect to x_i at the point $\mathbf{x}_0$. In this case, we have

$$\frac{\partial \mathbf{F}}{\partial x_i}(\mathbf{x}_0) = \left(\frac{\partial F_1}{\partial x_i}(\mathbf{x}_0), \ldots, \frac{\partial F_m}{\partial x_i}(\mathbf{x}_0) \right).$$

To capture all the partial derivatives, we define a derivative matrix.

Definition 4.5 The Derivative Matrix

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ be a function defined on $\mathcal{O}$. If $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ has partial derivative at the point $\mathbf{x}_0$, the derivative matrix of $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ at $\mathbf{x}_0$ is the $m \times n$ matrix

$$\mathbf{DF}(\mathbf{x}_0) = \begin{bmatrix} \nabla F_1(\mathbf{x}_0) \\ \nabla F_2(\mathbf{x}_0) \\ \vdots \\ \nabla F_m(\mathbf{x}_0) \end{bmatrix} = \begin{bmatrix} \dfrac{\partial F_1}{\partial x_1}(\mathbf{x}_0) & \dfrac{\partial F_1}{\partial x_2}(\mathbf{x}_0) & \cdots & \dfrac{\partial F_1}{\partial x_n}(\mathbf{x}_0) \\ \dfrac{\partial F_2}{\partial x_1}(\mathbf{x}_0) & \dfrac{\partial F_2}{\partial x_2}(\mathbf{x}_0) & \cdots & \dfrac{\partial F_2}{\partial x_n}(\mathbf{x}_0) \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial F_m}{\partial x_1}(\mathbf{x}_0) & \dfrac{\partial F_m}{\partial x_2}(\mathbf{x}_0) & \cdots & \dfrac{\partial F_m}{\partial x_n}(\mathbf{x}_0) \end{bmatrix}.$$

When $m = 1$, the derivative matrix is just the gradient of the function as a row matrix.

Example 4.5

Let $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^2$ be the function defined as

$$\mathbf{F}(x, y, z) = \left(xy^2z^3, x + 3y - 7z \right).$$

Find the derivative matrix of $\mathbf{F}$ at the point $(1, -1, 2)$.

Solution

$$\mathbf{DF}(x,y,z) = \begin{bmatrix} y^2z^3 & 2xyz^3 & 3xy^2z^2 \\ 1 & 3 & -7 \end{bmatrix}.$$

Thus, the derivative matrix of $\mathbf{F}$ at the point $(1,-1,2)$ is

$$\mathbf{DF}(1,-1,2) = \begin{bmatrix} 8 & -16 & 12 \\ 1 & 3 & -7 \end{bmatrix}.$$

Since the partial derivatives of a function is defined componentwise, we can focus on functions $f : \mathcal{O} \to \mathbb{R}$ whose codomain is $\mathbb{R}$. One might wonder why we have not mentioned the word "differentiable" so far. For single variable functions, we have seen in volume I that if a function is differentiable at a point, then it is continuous at that point. For multivariable functions, the existence of partial derivatives is not enough to guarantee continuity, as is shown in the next example.

Example 4.6

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function defined as

$$f(x,y) = \begin{cases} \dfrac{xy}{x^2+y^2}, & \text{if } (x,y) \neq (0,0), \\ 0, & \text{if } (x,y) = (0,0). \end{cases}$$

Show that f is not continuous at $(0,0)$, but it has partial derivatives at $(0,0)$.

Solution

Consider the sequence $\{\mathbf{u}_k\}$ with

$$\mathbf{u}_k = \left(\frac{1}{k}, \frac{1}{k}\right).$$

It is a sequence in $\mathbb{R}^2$ that converges to $(0,0)$. Since

$$f(\mathbf{u}_k) = \frac{1}{2} \qquad \text{for all } k \in \mathbb{Z}^+,$$

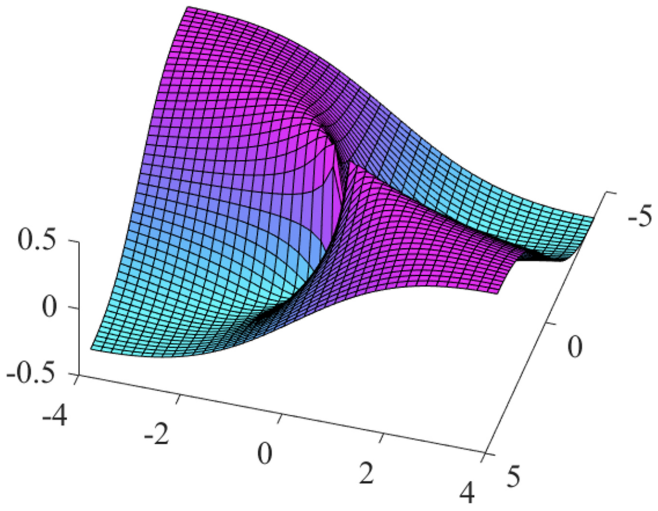


Figure 4.3: The function $f(x, y)$ defined in Example 4.6.

the sequence $\{f(\mathbf{u}_k)\}$ converges to $1/2$. But $f(0,0) = 0 \neq 1/2$. Since there is a sequence $\{\mathbf{u}_k\}$ that converges to $(0,0)$, but the sequence $\{f(\mathbf{u}_k)\}$ does not converge to $f(0,0)$, f is not continuous at $(0,0)$.
To find partial derivatives at $(0,0)$, we use definitions.

$$f_x(0,0) = \lim_{h\to 0} \frac{f(h,0) - f(0,0)}{h} = \lim_{h\to 0} \frac{0-0}{h} = 0,$$
$$f_y(0,0) = \lim_{h\to 0} \frac{f(0,h) - f(0,0)}{h} = \lim_{h\to 0} \frac{0-0}{h} = 0.$$

These show that f has partial derivatives at $(0,0)$, and $f_x(0,0) = f_y(0,0) = 0$.

For the function defined in Example 4.6, it has partial derivatives at all points. In fact, when $(x, y) \neq (0,0)$, we can apply derivative rules directly and find that

$$\frac{\partial f}{\partial x}(x,y) = \frac{(x^2+y^2)y - 2x^2y}{(x^2+y^2)^2} = \frac{y(y^2-x^2)}{(x^2+y^2)^2}.$$

Similarly,

$$\frac{\partial f}{\partial y}(x,y) = \frac{x(x^2-y^2)}{(x^2+y^2)^2}.$$

Let us highlight again our conclusion.

Partial Derivative vs Continuity

The existence of partial derivatives does not imply continuity.

This prompts us to find a better definition of *differentiability*, which can imply continuity. This will be considered in a latter section.

When the function $f : \mathcal{O} \to \mathbb{R}$ has partial derivative with respect to x_i, we obtain the function $f_{x_i} : \mathcal{O} \to \mathbb{R}$. Then we can discuss whether the function f_{x_i} has partial derivative at a point in $\mathcal{O}$.

Definition 4.6 Second Order Partial Derivatives

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $f : \mathcal{O} \to \mathbb{R}$ be a function defined on $\mathcal{O}$. Given that $1 \leq i \leq n$, $1 \leq j \leq n$, we say that the second order partial derivative $\dfrac{\partial^2 f}{\partial x_j \partial x_i}$ exists at $\mathbf{x}_0$ provided that there exists an open ball $B(\mathbf{x}_0, r)$ that is contained in $\mathcal{O}$ such that $\dfrac{\partial f}{\partial x_i} : B(\mathbf{x}_0, r) \to \mathbb{R}$ exists, and it has partial derivative with respect to x_j at the point $\mathbf{x}_0$. In this case, we define the second order partial derivative $\dfrac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0)$ of f at $\mathbf{x}_0$ as

$$\frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0) = \frac{\partial f_{x_i}}{\partial x_j}(\mathbf{x}_0) = \lim_{h \to 0} \frac{f_{x_i}(\mathbf{x}_0 + h\mathbf{e}_j) - f_{x_i}(\mathbf{x}_0)}{h}.$$

We say that the function $f : \mathcal{O} \to \mathbb{R}$ has second order partial derivatives at $\mathbf{x}_0$ provided that $\dfrac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0)$ exists for all $1 \leq i \leq n, 1 \leq j \leq n$.

In the same way, one can also define second order partial derivatives for a function $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ with codomain $\mathbb{R}^m$ when $m \geq 2$.

Remark 4.3

In the definition of the second order partial derivative $\dfrac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0)$, instead of assuming $f_{x_i}(\mathbf{x})$ exists for all $\mathbf{x}$ in a ball of radius r centered at $\mathbf{x}_0$, it is sufficient to assume that there exists $r > 0$ such that $f_{x_i}(\mathbf{x}_0 + h\mathbf{e}_j)$ exists for all $|h| < r$.

Definition 4.7

Given $1 \leq i \leq n, 1 \leq j \leq n$, we say that the function $f : \mathcal{O} \to \mathbb{R}$ has the second order partial derivative $\dfrac{\partial^2 f}{\partial x_j \partial x_i}$ provided that $\dfrac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0)$ exists for all $\mathbf{x}_0$ in $\mathcal{O}$.

We say that the function $f : \mathcal{O} \to \mathbb{R}$ has second order partial derivatives provided that $\dfrac{\partial^2 f}{\partial x_j \partial x_i}$ exists for all $1 \leq i \leq n, 1 \leq j \leq n$.

Notations for Second Order Partial Derivatives

Alternative notations for second order partial derivatives are

$$\frac{\partial^2 f}{\partial x_j \partial x_i} = (f_{x_i})_{x_j} = f_{x_i x_j}.$$

Notice that the orders of x_i and x_j are different in different notations.

Remark 4.4

Given $1 \leq i \leq n, 1 \leq j \leq n$, the function $f : \mathcal{O} \to \mathbb{R}$ has the second order partial derivative $\dfrac{\partial^2 f}{\partial x_j \partial x_i}$ provided that $f_{x_i} : \mathcal{O} \to \mathbb{R}$ exists, and f_{x_i} has partial derivative with respect to x_j.

Example 4.7

Find the second order partial derivatives of the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined as

$$f(x, y) = xe^{2x+3y}.$$

Solution

We find the first order partial derivatives first.

$$\frac{\partial f}{\partial x}(x,y) = e^{2x+3y} + 2xe^{2x+3y} = (1+2x)e^{2x+3y},$$
$$\frac{\partial f}{\partial y}(x,y) = 3xe^{2x+3y}.$$

Then we compute the second order partial derivatives.

$$\frac{\partial^2 f}{\partial x^2}(x,y) = 2e^{2x+3y} + 2(1+2x)e^{2x+3y} = (4+4x)e^{2x+3y},$$
$$\frac{\partial^2 f}{\partial y \partial x}(x,y) = 3(1+2x)e^{2x+3y} = (3+6x)e^{2x+3y},$$
$$\frac{\partial^2 f}{\partial x \partial y}(x,y) = 3e^{2x+3y} + 6xe^{2x+3y} = (3+6x)e^{2x+3y},$$
$$\frac{\partial^2 f}{\partial y^2}(x,y) = 9xe^{2x+3y}.$$

Definition 4.8 The Hessian Matrix

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$. If $f : \mathcal{O} \to \mathbb{R}$ is a function that has second order partial derivatives at $\mathbf{x}_0$, the Hessian matrix of f at $\mathbf{x}_0$ is the $n \times n$ matrix defined as

$$H_f(\mathbf{x}_0) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}_0)\right]$$
$$= \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(\mathbf{x}_0) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(\mathbf{x}_0) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\mathbf{x}_0) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(\mathbf{x}_0) & \frac{\partial^2 f}{\partial x_2^2}(\mathbf{x}_0) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(\mathbf{x}_0) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\mathbf{x}_0) & \frac{\partial^2 f}{\partial x_n \partial x_2}(\mathbf{x}_0) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(\mathbf{x}_0) \end{bmatrix}.$$

We do not define Hessian matrix for a function $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ with codomain $\mathbb{R}^m$ when $m \geq 2$.

Example 4.8

For the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined as $f(x, y) = xe^{2x+3y}$ in Example 4.7,

$$H_f(x, y) = \begin{bmatrix} (4 + 4x)e^{2x+3y} & (3 + 6x)e^{2x+3y} \\ (3 + 6x)e^{2x+3y} & 9xe^{2x+3y} \end{bmatrix}.$$

In Example 4.7, we notice that

$$\frac{\partial^2 f}{\partial y \partial x}(x, y) = \frac{\partial^2 f}{\partial x \partial y}(x, y)$$

for all $(x, y) \in \mathbb{R}^2$. The following example shows that *this is not always true*.

Example 4.9

Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined as

$$f(x, y) = \begin{cases} \dfrac{xy(x^2 - y^2)}{x^2 + y^2}, & \text{if } (x, y) \neq (0, 0), \\ 0, & \text{if } (x, y) = (0, 0). \end{cases}$$

Find $f_{xy}(0, 0)$ and $f_{yx}(0, 0)$.

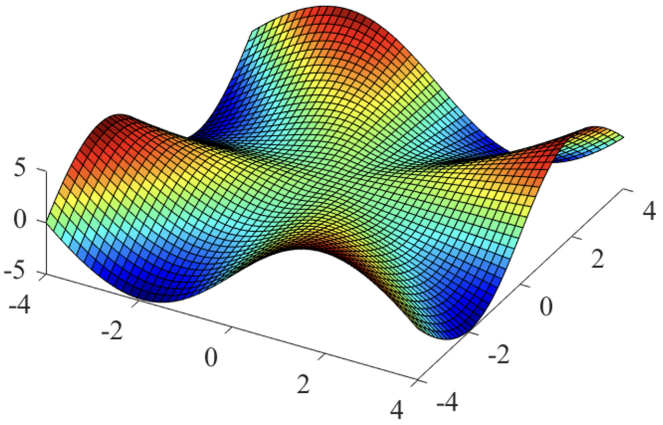


Figure 4.4: The function $f(x, y)$ defined in Example 4.9.

Solution

To compute $f_{xy}(0,0)$, we need to compute $f_x(0,h)$ for all h in a neighbourhood of 0. To compute $f_{yx}(0,0)$, we need to compute $f_y(h,0)$ for all h in a neighbourhood of 0. Notice that for any $h \in \mathbb{R}$, $f(0,h) = f(h,0) = 0$. By considering $h = 0$ and $h \neq 0$ separately, we find that

$$f_x(0,h) = \lim_{t\to 0} \frac{f(t,h) - f(0,h)}{t} = \lim_{t\to 0} \frac{h(t^2-h^2)}{t^2+h^2} = -h,$$
$$f_y(h,0) = \lim_{t\to 0} \frac{f(h,t) - f(h,0)}{t} = \lim_{t\to 0} \frac{h(h^2-t^2)}{h^2+t^2} = h.$$

It follows that

$$f_{xy}(0,0) = \lim_{h\to 0} \frac{f_x(0,h) - f_x(0,0)}{h} = \lim_{h\to 0} \frac{-h}{h} = -1,$$
$$f_{yx}(0,0) = \lim_{h\to 0} \frac{f_y(h,0) - f_y(0,0)}{h} = \lim_{h\to 0} \frac{h}{h} = 1.$$

Example 4.9 shows that there exists a function $f : \mathbb{R}^2 \to \mathbb{R}$ which has second order partial derivatives at $(0,0)$ but

$$\frac{\partial^2 f}{\partial x \partial y}(0,0) \neq \frac{\partial^2 f}{\partial y \partial x}(0,0).$$

Remark 4.5

If $\mathcal{O}$ is an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, there exists $r > 0$ such that $B(\mathbf{x}_0, r) \subset \mathcal{O}$. Given that $f : \mathcal{O} \to \mathbb{R}$ is a function defined on $\mathcal{O}$, and $1 \leq i < j \leq n$, let $\mathfrak{D}$ be the ball with center at $(0,0)$ and radius r in $\mathbb{R}^2$. Define the function $g : \mathfrak{D} \to \mathbb{R}$ by

$$g(u,v) = f(\mathbf{x}_0 + u\mathbf{e}_i + v\mathbf{e}_j).$$

Then $\dfrac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0)$ exists if and only if $\dfrac{\partial^2 g}{\partial v \partial u}(0,0)$ exists. In such case, we have

$$\frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0) = \frac{\partial^2 g}{\partial v \partial u}(0,0).$$

The following gives a sufficient condition to interchange the order of taking partial derivatives.

Theorem 4.2 Clairaut's Theorem or Schwarz's Theorem

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $f : \mathcal{O} \to \mathbb{R}$ be a function defined on $\mathcal{O}$. Assume that $1 \leq i < j \leq n$, and the second order partial derivatives $\dfrac{\partial^2 f}{\partial x_j \partial x_i} : \mathcal{O} \to \mathbb{R}$ and $\dfrac{\partial^2 f}{\partial x_i \partial x_j} : \mathcal{O} \to \mathbb{R}$ exist. If the functions $\dfrac{\partial^2 f}{\partial x_j \partial x_i}$ and $\dfrac{\partial^2 f}{\partial x_i \partial x_j} : \mathcal{O} \to \mathbb{R}$ are continuous at $\mathbf{x}_0$, then

$$\frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0) = \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}_0).$$

Proof

Since $\mathcal{O}$ is an open set that contains the point $\mathbf{x}_0$, there exists $r > 0$ such that $B(\mathbf{x}_0, r) \subset \mathcal{O}$. Let

$$\mathfrak{D} = \left\{(u, v) \middle| \, u^2 + v^2 < r^2\right\},$$

and define the function $g : \mathfrak{D} \to \mathbb{R}$ by

$$g(u, v) = f(\mathbf{x}_0 + u\mathbf{e}_i + v\mathbf{e}_j).$$

By Remark 4.5, g has second order partial derivatives, and $\dfrac{\partial^2 g}{\partial v \partial u}$ and $\dfrac{\partial^2 g}{\partial u \partial v}$ are continuous at $(0, 0)$. We need to show that

$$\frac{\partial^2 g}{\partial v \partial u}(0, 0) = \frac{\partial^2 g}{\partial u \partial v}(0, 0).$$

Consider the function

$$G(u, v) = g(u, v) - g(u, 0) - g(0, v) + g(0, 0).$$

Notice that

$$G(u, v) = H_v(u) - H_v(0) = S_u(v) - S_u(0),$$

where

$$H_v(u) = g(u,v) - g(u,0), \qquad S_u(v) = g(u,v) - g(0,v).$$

For fixed v with $|v| < r$, the function $H_v(u)$ is defined for those u with $|u| < \sqrt{r^2 - v^2}$, such that (u,v) is in $\mathfrak{D}$. It is differentiable with

$$H_v'(u) = \frac{\partial g}{\partial u}(u,v) - \frac{\partial g}{\partial u}(u,0).$$

Hence, if (u,v) is in $\mathfrak{D}$, mean value theorem for single variable functions implies that there exists $c_{u,v} \in (0,1)$ such that

$$\begin{aligned} G(u,v) &= H_v(u) - H_v(0) \\ &= uH_v'(c_{u,v}u) \\ &= u\left(\frac{\partial g}{\partial u}(c_{u,v}u, v) - \frac{\partial g}{\partial u}(c_{u,v}u, 0)\right). \end{aligned}$$

Regard this now as a function of v, the mean value theorem for single variable functions implies that there exists $d_{u,v} \in (0,1)$ such that

$$G(u,v) = uv\frac{\partial^2 g}{\partial v \partial u}(c_{u,v}u, d_{u,v}v). \tag{4.1}$$

Using the same reasoning, we find that for $(u,v) \in \mathfrak{D}$, there exists $\widetilde{d}_{u,v} \in (0,1)$ such that

$$G(u,v) = vS_u'(\widetilde{d}_{u,v}v) = v\left(\frac{\partial g}{\partial v}(u, \widetilde{d}_{u,v}v) - \frac{\partial g}{\partial v}(0, \widetilde{d}_{u,v}v)\right).$$

Regard this as a function of u, mean value theorem implies that there exists $\widetilde{c}_{u,v} \in (0,1)$ such that

$$G(u,v) = uv\frac{\partial^2 g}{\partial u \partial v}(\widetilde{c}_{u,v}u, \widetilde{d}_{u,v}v). \tag{4.2}$$

Comparing (4.1) and (4.2), we find that

$$\frac{\partial^2 g}{\partial v \partial u}(c_{u,v}u, d_{u,v}v) = \frac{\partial^2 g}{\partial u \partial v}(\widetilde{c}_{u,v}u, \widetilde{d}_{u,v}v).$$

When $(u, v) \to (0, 0)$, $(c_{u,v}u, d_{u,v}v) \to (0, 0)$ and $(\widetilde{c}_{u,v}u, \widetilde{d}_{u,v}v) \to (0, 0)$. The continuities of g_{uv} and g_{vu} at $(0, 0)$ then imply that

$$\frac{\partial^2 g}{\partial v \partial u}(0, 0) = \frac{\partial^2 g}{\partial u \partial v}(0, 0).$$

This completes the proof.

Example 4.10

Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ in Example 4.9 defined as

$$f(x, y) = \begin{cases} \dfrac{xy(x^2 - y^2)}{x^2 + y^2}, & \text{if } (x, y) \neq (0, 0), \\ 0, & \text{if } (x, y) = (0, 0). \end{cases}$$

When $(x, y) \neq (0, 0)$, we find that

$$\frac{\partial f}{\partial x}(x, y) = \frac{y(x^4 + 4x^2y^2 - y^4)}{(x^2 + y^2)^2},$$
$$\frac{\partial f}{\partial y}(x, y) = \frac{x(x^4 - 4x^2y^2 - y^4)}{(x^2 + y^2)^2}.$$

It follows that

$$\frac{\partial^2 f}{\partial y \partial x}(x, y) = \frac{x^6 + 9x^4y^2 - 9x^2y^4 - y^6}{(x^2 + y^2)^3} = \frac{\partial^2 f}{\partial x \partial y}(x, y).$$

Indeed, both f_{xy} and f_{yx} are continuous on $\mathbb{R}^2 \setminus \{(0, 0)\}$.

Corollary 4.3

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $f : \mathcal{O} \to \mathbb{R}$ be a function defined on $\mathcal{O}$. If all the second order partial derivatives of the function $f : \mathcal{O} \to \mathbb{R}$ at $\mathbf{x}_0$ are continuous, then the Hessian matrix $H_f(\mathbf{x}_0)$ of f at $\mathbf{x}_0$ is a symmetric matrix.

Remark 4.6

One can define partial derivatives of higher orders following the same rationale as we define the second order partial derivatives. Extension of Clairaut's theorem to higher order partial derivatives is straightforward. The key point is the continuity of the partial derivatives involved.

Exercises 4.1

Question 1

Let $f : \mathbb{R}^3 \to \mathbb{R}$ be the function defined as

$$f(x, y, z) = \frac{xz}{e^y + 1}.$$

Find $\nabla f(1, 0, -1)$, the gradient of f at the point $(1, 0, -1)$.

Question 2

Let $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^3$ be the function defined as

$$\mathbf{F}(x, y) = \left(x^2 y, xy^2, 3x^2 + 4y^2\right).$$

Find $\mathbf{DF}(2, -1)$, the derivative matrix of $\mathbf{F}$ at the point $(2, -1)$.

Question 3

Let $f : \mathbb{R}^3 \to \mathbb{R}$ be the function defined as

$$f(x, y, z) = x^2 + 3xyz + 2y^2 z^3.$$

Find $H_f(1, -1, 2)$, the Hessian matrix of f at the point $(1, -1, 2)$.

Question 4

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function defined as

$$f(x, y) = \begin{cases} \dfrac{3xy}{x^2 + 4y^2}, & \text{if } (x, y) \neq (0, 0), \\ 0, & \text{if } (x, y) = (0, 0). \end{cases}$$

Show that f is not continuous at $(0, 0)$, but it has partial derivatives at $(0, 0)$.

Question 5

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function defined as $f(x, y) = |x^2 + y|$. Determine whether $f_y(1, -1)$ exists.

Question 6

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function defined as

$$f(x,y) = \begin{cases} \dfrac{x^2 y}{x^2+y^2}, & \text{if } (x,y) \neq (0,0), \\ 0, & \text{if } (x,y) = (0,0). \end{cases}$$

Show that f is continuous, it has partial derivatives, but the partial derivatives are not continuous.

Question 7

Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined as

$$f(x,y) = \begin{cases} \dfrac{xy(x^2+9y^2)}{4x^2+y^2}, & \text{if } (x,y) \neq (0,0), \\ 0, & \text{if } (x,y) = (0,0). \end{cases}$$

Find the Hessian matrix $H_f(0,0)$ of f at $(0,0)$.

4.2 Differentiability and First Order Approximation

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ be a function defined on $\mathcal{O}$. As we have seen in the previous section, even if $\mathbf{F}$ has partial derivatives at $\mathbf{x}_0$, it does not imply that $\mathbf{F}$ is continuous at $\mathbf{x}_0$. Heuristically, this is because the partial derivatives only consider the change of the function along the n directions defined by the coordinate axes, while continuity of $\mathbf{F}$ requires us to consider the change of $\mathbf{F}$ along *all* directions.

4.2.1 Differentiability

In this section, we will give a suitable definition of *differentiability* to ensure that we can capture the change of $\mathbf{F}$ in all directions. Let us first revisit an alternative perpective of *differentiability* for a single variable function $f : (a, b) \to \mathbb{R}$, which we have discussed in volume I. If x_0 is a point in (a, b), then the function $f : (a, b) \to \mathbb{R}$ is differentiable at x_0 if and only if there is a number c such that

$$\lim_{h \to 0} \frac{f(x_0 + h) - f(x_0) - ch}{h} = 0. \tag{4.3}$$

In fact, if f is differentiable at x_0, then this number c has to equal to $f'(x_0)$.

Now for a function $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ defined on an open subset $\mathcal{O}$ of $\mathbb{R}^n$, to consider the differentiability of $\mathbf{F}$ at $\mathbf{x}_0 \in \mathcal{O}$, we should compare $\mathbf{F}(\mathbf{x}_0)$ to $\mathbf{F}(\mathbf{x}_0 + \mathbf{h})$ for all $\mathbf{h}$ in a neighbourhood of $\mathbf{0}$. But then a reasonable substitute of the number c should be a linear transformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^m$, so that for each $\mathbf{h}$ in a neighbourhood of $\mathbf{0}$, it gives a vector $\mathbf{T}(\mathbf{h})$ in $\mathbb{R}^m$. As now $\mathbf{h}$ is a vector in $\mathbb{R}^n$, we cannot divide by $\mathbf{h}$ in (4.3). It should be replaced with $\|\mathbf{h}\|$, the norm of $\mathbf{h}$.

Definition 4.9 Differentiability

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ be a function defined on $\mathcal{O}$. The function $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is *differentiable* at $\mathbf{x}_0$ provided that there exists a linear transformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^m$ so that

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{\mathbf{F}(\mathbf{x}_0 + \mathbf{h}) - \mathbf{F}(\mathbf{x}_0) - \mathbf{T}(\mathbf{h})}{\|\mathbf{h}\|} = \mathbf{0}.$$

$\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is differentiable if it is differentiable at each point of $\mathcal{O}$.

Remark 4.7

The differentiability of $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ at $\mathbf{x}_0$ amounts to the existence of a linear transformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^m$ so that

$$\mathbf{F}(\mathbf{x}_0 + \mathbf{h}) = \mathbf{F}(\mathbf{x}_0) + \mathbf{T}(\mathbf{h}) + \boldsymbol{\varepsilon}(\mathbf{h})\|\mathbf{h}\|,$$

where $\boldsymbol{\varepsilon}(\mathbf{h}) \to \mathbf{0}$ as $\mathbf{h} \to \mathbf{0}$.

The following is obvious from the definition.

Proposition 4.4

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ be a function defined on $\mathcal{O}$. The function $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is differentiable at $\mathbf{x}_0$ if and only if each of its component functions $F_j : \mathcal{O} \to \mathbb{R}$, $1 \le j \le m$ is differentiable at $\mathbf{x}_0$.

Proof

Let the components of the function

$$\boldsymbol{\varepsilon}(\mathbf{h}) = \frac{\mathbf{F}(\mathbf{x}_0 + \mathbf{h}) - \mathbf{F}(\mathbf{x}_0) - \mathbf{T}(\mathbf{h})}{\|\mathbf{h}\|}$$

be $\varepsilon_1(\mathbf{h}), \varepsilon_2(\mathbf{h}), \ldots, \varepsilon_m(\mathbf{h})$. Then for $1 \le j \le m$,

$$\varepsilon_j(\mathbf{h}) = \frac{F_j(\mathbf{x}_0 + \mathbf{h}) - F_j(\mathbf{x}_0) - T_j(\mathbf{h})}{\|\mathbf{h}\|}.$$

The assertion of the proposition follows from the fact that

$$\lim_{\mathbf{h}\to\mathbf{0}} \boldsymbol{\varepsilon}(\mathbf{h}) = \mathbf{0} \quad \text{if and only if} \quad \lim_{\mathbf{h}\to\mathbf{0}} \varepsilon_j(\mathbf{h}) = 0 \quad \text{for all} \quad 1 \le j \le m,$$

while $\lim_{\mathbf{h}\to\mathbf{0}} \varepsilon_j(\mathbf{h}) = 0$ if and only if $F_j : \mathcal{O} \to \mathbb{R}$ is differentiable at $\mathbf{x}_0$.

Let us look at a simple example of differentiable functions.

Example 4.11

Let A be an $m \times n$ matrix, and let $\mathbf{b}$ be a point in $\mathbb{R}^m$. Define the function $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$ by

$$\mathbf{F}(\mathbf{x}) = A\mathbf{x} + \mathbf{b}.$$

Show that $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable.

Solution

Given $\mathbf{x}_0$ and $\mathbf{h}$ in $\mathbb{R}^n$, notice that

$$\mathbf{F}(\mathbf{x}_0 + \mathbf{h}) - \mathbf{F}(\mathbf{x}_0) = A(\mathbf{x}_0 + \mathbf{h}) + \mathbf{b} - A\mathbf{x}_0 - \mathbf{b} = A\mathbf{h}. \tag{4.4}$$

The map $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^m$ defined as $\mathbf{T}(\mathbf{h}) = A\mathbf{h}$ is a linear transformation. Eq. (4.4) says that

$$\mathbf{F}(\mathbf{x}_0 + \mathbf{h}) - \mathbf{F}(\mathbf{x}_0) - \mathbf{T}(\mathbf{h}) = \mathbf{0}.$$

Thus,

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\mathbf{F}(\mathbf{x}_0 + \mathbf{h}) - \mathbf{F}(\mathbf{x}_0) - \mathbf{T}(\mathbf{h})}{\|\mathbf{h}\|} = \mathbf{0}.$$

Therefore, $\mathbf{F}$ is differentiable at $\mathbf{x}_0$. Since the point $\mathbf{x}_0$ is arbitrary, the function $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable.

The next theorem says that differentiability implies continuity.

Theorem 4.5 Differentiability Implies Continuity

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ be a function defined on $\mathcal{O}$. If the function $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is differentiable at $\mathbf{x}_0$, then it is continuous at $\mathbf{x}_0$.

Proof

Since $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is differentiable at $\mathbf{x}_0$, there exists a linear transformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$\boldsymbol{\varepsilon}(\mathbf{h}) = \frac{\mathbf{F}(\mathbf{x}_0 + \mathbf{h}) - \mathbf{F}(\mathbf{x}_0) - \mathbf{T}(\mathbf{h})}{\|\mathbf{h}\|} \xrightarrow{\mathbf{h}\to\mathbf{0}} \mathbf{0}.$$

By Theorem 2.34, there is a positive constant c such that

$$\|\mathbf{T}(\mathbf{h})\| \leq c\|\mathbf{h}\| \qquad \text{for all } \mathbf{h} \in \mathbb{R}^n.$$

Therefore,

$$\|\mathbf{F}(\mathbf{x}_0+\mathbf{h}) - \mathbf{F}(\mathbf{x}_0)\| \leq \|\mathbf{T}(\mathbf{h})\| + \|\mathbf{h}\|\|\boldsymbol{\varepsilon}(\mathbf{h})\| \leq \|\mathbf{h}\|\left(c + \|\boldsymbol{\varepsilon}(\mathbf{h})\|\right).$$

This implies that

$$\lim_{\mathbf{h}\to\mathbf{0}} \mathbf{F}(\mathbf{x}_0+\mathbf{h}) = \mathbf{F}(\mathbf{x}_0).$$

Thus, $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is continuous at $\mathbf{x}_0$.

Example 4.12

The function $f : \mathbb{R}^2 \to \mathbb{R}$ defined as

$$f(x,y) = \begin{cases} \dfrac{xy}{x^2+y^2}, & \text{if } (x,y) \neq (0,0), \\ 0, & \text{if } (x,y) = (0,0) \end{cases}$$

in Example 4.6 is not differentiable at $(0,0)$ since it is not continuous at $(0,0)$. However, we have shown that it has partial derivatives at $(0,0)$.

Let us study the function $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{F}(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$ that is defined in Example 4.11. The component functions of $\mathbf{F}$ are

$$\begin{aligned} F_1(x_1, x_2, \ldots, x_n) &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n + b_1, \\ F_2(x_1, x_2, \ldots, x_n) &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n + b_2, \\ &\vdots \\ F_m(x_1, x_2, \ldots, x_n) &= a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n + b_m. \end{aligned}$$

Notice that

$$\begin{aligned} \nabla F_1(\mathbf{x}) &= \mathbf{a}_1 = (a_{11}, a_{12}, \ldots, a_{1n}), \\ \nabla F_2(\mathbf{x}) &= \mathbf{a}_2 = (a_{21}, a_{22}, \ldots, a_{2n}), \\ &\vdots \\ \nabla F_m(\mathbf{x}) &= \mathbf{a}_m = (a_{m1}, a_{m2}, \ldots, a_{mn}) \end{aligned}$$

are the row vectors of A. Hence, the derivative matrix of $\mathbf{F}$ is a given by

$$\mathbf{DF}(\mathbf{x}) = \begin{bmatrix} \nabla F_1(\mathbf{x}) \\ \nabla F_2(\mathbf{x}) \\ \vdots \\ \nabla F_m(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix},$$

which is the matrix A itself. Observe that

$$\mathrm{D}\mathbf{F}(\mathbf{x})\mathbf{h} = \begin{bmatrix} a_{11}h_1 + a_{12}h_2 + \cdots + a_{1n}h_n \\ a_{21}h_1 + a_{22}h_2 + \cdots + a_{2n}h_n \\ \vdots \\ a_{m1}h_1 + a_{m2}h_2 + \cdots + a_{mn}h_n \end{bmatrix} = \begin{bmatrix} \langle \nabla F_1(\mathbf{x}), \mathbf{h} \rangle \\ \langle \nabla F_2(\mathbf{x}), \mathbf{h} \rangle \\ \vdots \\ \langle \nabla F_m(\mathbf{x}), \mathbf{h} \rangle \end{bmatrix}.$$

From Example 4.11, we suspect that the linear transformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^m$ that appears in the definition of differentiability of a function should be the linear transformation defined by the derivative matrix. In fact, this is the case.

Theorem 4.6

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ be a function defined on $\mathcal{O}$. The following are equivalent.

(a) The function $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is differentiable at $\mathbf{x}_0$.

(b) The function $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ has partial derivatives at $\mathbf{x}_0$, and

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{\mathbf{F}(\mathbf{x}_0 + \mathbf{h}) - \mathbf{F}(\mathbf{x}_0) - \mathbf{DF}(\mathbf{x}_0)\mathbf{h}}{\|\mathbf{h}\|} = \mathbf{0}. \tag{4.5}$$

(c) For each $1 \leq j \leq m$, the component function $F_j : \mathcal{O} \to \mathbb{R}$ has partial derivatives at $\mathbf{x}_0$, and

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{F_j(\mathbf{x}_0 + \mathbf{h}) - F_j(\mathbf{x}_0) - \langle \nabla F_j(\mathbf{x}_0), \mathbf{h} \rangle}{\|\mathbf{h}\|} = 0.$$

Proof

The equivalence of (b) and (c) is Proposition 4.4, the componentwise differentiability. Thus, we are left to prove the equivalence of (a) and (b).
First, we prove (b) implies (a). If (b) holds, let $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^m$ be the linear transformation defined by the derivative matrix $\mathbf{DF}(\mathbf{x}_0)$. Then (4.5) says that $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is differentiable at $\mathbf{x}_0$.
Conversely, assume that $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is differentiable at $\mathbf{x}_0$. Then there exists a linear transformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\mathbf{F}(\mathbf{x}_0+\mathbf{h}) - \mathbf{F}(\mathbf{x}_0) - \mathbf{T}(\mathbf{h})}{\|\mathbf{h}\|} = \mathbf{0}. \tag{4.6}$$

Let A be a $m \times n$ matrix so that $\mathbf{T}(\mathbf{h}) = A\mathbf{h}$. For $1 \le i \le n$, eq. (4.6) implies that

$$\lim_{h\to 0} \frac{\mathbf{F}(\mathbf{x}_0+h\mathbf{e}_i) - \mathbf{F}(\mathbf{x}_0) - A(h\mathbf{e}_i)}{h} = \mathbf{0}.$$

This gives

$$A\mathbf{e}_i = \lim_{h\to 0} \frac{\mathbf{F}(\mathbf{x}_0+h\mathbf{e}_i) - \mathbf{F}(\mathbf{x}_0)}{h}.$$

This shows that $\dfrac{\partial \mathbf{F}}{\partial x_i}(\mathbf{x}_0)$ exists and

$$\frac{\partial \mathbf{F}}{\partial x_i}(\mathbf{x}_0) = A\mathbf{e}_i.$$

Therefore, $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ has partial derivatives at $\mathbf{x}_0$. Since

$$\begin{aligned} A &= \left[A\mathbf{e}_1 \;\middle|\; A\mathbf{e}_2 \;\middle|\; \cdots \;\middle|\; A\mathbf{e}_n\right] \\ &= \left[\frac{\partial \mathbf{F}}{\partial x_1}(\mathbf{x}_0) \;\middle|\; \frac{\partial \mathbf{F}}{\partial x_2}(\mathbf{x}_0) \;\middle|\; \cdots \;\middle|\; \frac{\partial \mathbf{F}}{\partial x_n}(\mathbf{x}_0)\right] = \mathbf{DF}(\mathbf{x}_0), \end{aligned}$$

eq. (4.6) says that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\mathbf{F}(\mathbf{x}_0+\mathbf{h}) - \mathbf{F}(\mathbf{x}_0) - \mathbf{DF}(\mathbf{x}_0)\mathbf{h}}{\|\mathbf{h}\|} = \mathbf{0}.$$

This proves (a) implies (b).

Corollary 4.7

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ be a function defined on $\mathcal{O}$. If the partial derivatives of $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ exist at $\mathbf{x}_0$, but

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\mathbf{F}(\mathbf{x}_0+\mathbf{h}) - \mathbf{F}(\mathbf{x}_0) - \mathbf{DF}(\mathbf{x}_0)\mathbf{h}}{\|\mathbf{h}\|} \neq \mathbf{0},$$

then $\mathbf{F}$ is not differentiable at $\mathbf{x}_0$.

Proof

If $\mathbf{F}$ is differentiable at $\mathbf{x}_0$, Theorem 4.6 says that we must have

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\mathbf{F}(\mathbf{x}_0+\mathbf{h}) - \mathbf{F}(\mathbf{x}_0) - \mathbf{DF}(\mathbf{x}_0)\mathbf{h}}{\|\mathbf{h}\|} = \mathbf{0}.$$

By contrapositive, since

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\mathbf{F}(\mathbf{x}_0+\mathbf{h}) - \mathbf{F}(\mathbf{x}_0) - \mathbf{DF}(\mathbf{x}_0)\mathbf{h}}{\|\mathbf{h}\|} \neq \mathbf{0},$$

we find that $\mathbf{F}$ is not differentiable at $\mathbf{x}_0$.

Example 4.13

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function defined as

$$f(x,y) = \begin{cases} \dfrac{x^3}{x^2+y^2}, & \text{if } (x,y) \neq (0,0), \\ 0, & \text{if } (x,y) = (0,0). \end{cases}$$

Determine whether f is differentiable at $(0,0)$.

Solution

One can show that f is continuous at $\mathbf{0} = (0,0)$. Hence, we cannot use continuity to determine whether f is differentiable at $\mathbf{x}_0$. Notice that

$$f_x(0,0) = \lim_{h\to 0} \frac{f(h,0) - f(0,0)}{h} = \lim_{h\to 0} \frac{h-0}{h} = 1,$$

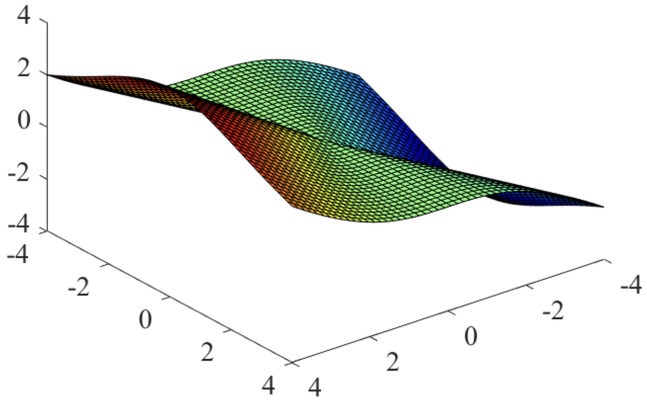


Figure 4.5: The function $f(x, y)$ defined in Example 4.13.

$$f_y(0,0) = \lim_{h\to 0} \frac{f(0,h) - f(0,0)}{h} = \lim_{h\to 0} \frac{0-0}{h} = 0.$$

Therefore, f has partial derivatives at $\mathbf{0}$, and $\nabla f(\mathbf{0}) = (1, 0)$. Now we consider the function

$$\varepsilon(\mathbf{h}) = \frac{f(\mathbf{h}) - f(\mathbf{0}) - \langle \nabla f(\mathbf{0}), \mathbf{h}\rangle}{\|\mathbf{h}\|} = -\frac{h_1 h_2^2}{(h_1^2 + h_2^2)^{3/2}}.$$

Let $\{\mathbf{h}_k\}$ be the sequence with $\mathbf{h}_k = \left(\frac{1}{k}, \frac{1}{k}\right)$. It converges to $\mathbf{0}$. Since

$$\varepsilon(\mathbf{h}_k) = -\frac{1}{2\sqrt{2}} \qquad \text{for all } k \in \mathbb{Z}^+,$$

The sequence $\{\varepsilon(\mathbf{h}_k)\}$ does not converge to 0. Hence,

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{f(\mathbf{h}) - f(\mathbf{0}) - \langle \nabla f(\mathbf{0}), \mathbf{h}\rangle}{\|\mathbf{h}\|} \neq 0.$$

Therefore, f is not differentiable at $(0, 0)$.

Example 4.13 gives a function which is continuous and has partial derivatives at a point, yet it fails to be differentiable at that point. In the following, we are going to give a sufficient condition for differentiability. We begin with a lemma.

Lemma 4.8

Let $\mathbf{x}_0$ be a point in $\mathbb{R}^n$ and let $f : B(\mathbf{x}_0, r) \to \mathbb{R}$ be a function defined on an open ball centered at $\mathbf{x}_0$. Assume that $f : B(\mathbf{x}_0, r) \to \mathbb{R}$ has first order partial derivatives. For each $\mathbf{h}$ in $\mathbb{R}^n$ with $\|\mathbf{h}\| < r$, there exists $\mathbf{z}_1, \ldots, \mathbf{z}_n$ in $B(\mathbf{x}_0, r)$ such that

$$f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) = \sum_{i=1}^{n} h_i \frac{\partial f}{\partial x_i}(\mathbf{z}_i),$$

and

$$\|\mathbf{z}_i - \mathbf{x}_0\| < \|\mathbf{h}\| \qquad \text{for all } 1 \leq i \leq n.$$

Proof

We will take a zigzag path from $\mathbf{x}_0$ to $\mathbf{x}_0 + \mathbf{h}$, which is a union of paths parallel to the coordinate axes. For $1 \leq i \leq n$, let

$$\mathbf{x}_i = \mathbf{x}_0 + \sum_{k=1}^{i} h_k \mathbf{e}_k = \mathbf{x}_0 + h_1 \mathbf{e}_1 + \cdots + h_i \mathbf{e}_i.$$

Then $\mathbf{x}_i$ is in $B(\mathbf{x}_0, r)$. Notice that $B(\mathbf{x}_0, r)$ is a convex set. Therefore, for any $1 \leq i \leq n$, the line segment between $\mathbf{x}_{i-1}$ and $\mathbf{x}_i = \mathbf{x}_{i-1} + h_i \mathbf{e}_i$ lies entirely inside $B(\mathbf{x}_0, r)$. Since $f : B(\mathbf{x}_0, r) \to \mathbb{R}$ has first order partial derivative with respect to x_i, the function $g_i : [0, 1] \to \mathbb{R}$,

$$g_i(t) = f(\mathbf{x}_{i-1} + t h_i \mathbf{e}_i)$$

is differentiable and

$$g_i'(t) = h_i \frac{\partial f}{\partial x_i}(\mathbf{x}_{i-1} + t h_i \mathbf{e}_i).$$

By mean value theorem, there exists $c_i \in (0, 1)$ such that

$$f(\mathbf{x}_i) - f(\mathbf{x}_{i-1}) = g_i(1) - g_i(0) = g_i'(c_i) = h_i \frac{\partial f}{\partial x_i}(\mathbf{x}_{i-1} + c_i h_i \mathbf{e}_i).$$

Let

$$\mathbf{z}_i = \mathbf{x}_{i-1} + c_i h_i \mathbf{e}_i = \mathbf{x}_0 + \sum_{k=1}^{i-1} h_k \mathbf{e}_k + c_i h_i \mathbf{e}_i.$$

Then $\mathbf{z}_i$ is a point in $B(\mathbf{x}_0, r)$. Moreover,

$$f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) = \sum_{i=1}^{n} (f(\mathbf{x}_i) - f(\mathbf{x}_{i-1})) = \sum_{i=1}^{n} h_i \frac{\partial f}{\partial x_i}(\mathbf{z}_i).$$

For $1 \leq i \leq n$, since $c_i \in (0, 1)$, we have

$$\|\mathbf{z}_i - \mathbf{x}_0\| = \sqrt{h_1^2 + \cdots + h_{i-1}^2 + c_i^2 h_i^2} < \sqrt{h_1^2 + \cdots + h_{i-1}^2 + h_i^2} \leq \|\mathbf{h}\|.$$

This completes the proof.

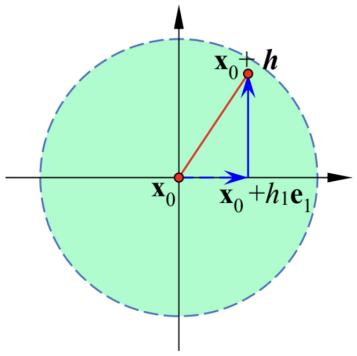


Figure 4.6: A zigzag path from $\mathbf{x}_0$ to $\mathbf{x}_0 + \mathbf{h}$.

Theorem 4.9

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ be a function defined on $\mathcal{O}$. If the partial derivatives of $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ exists and are continuous at $\mathbf{x}_0$, then $\mathbf{F}$ is differentiable at $\mathbf{x}_0$.

Proof

By Proposition 4.4, it suffices to prove the theorem for a function $f : \mathcal{O} \to \mathbb{R}$ with codomain $\mathbb{R}$. Since $\mathcal{O}$ is an open set that contains the point $\mathbf{x}_0$, there exists $r > 0$ such that $B(\mathbf{x}_0, r) \subset \mathcal{O}$. By Lemma 4.8, for each $\mathbf{h}$ that satisfies $0 < \|\mathbf{h}\| < r$, there exists $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ such that

$$f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) = \sum_{i=1}^{n} h_i \frac{\partial f}{\partial x_i}(\mathbf{z}_i),$$

and

$$\|\mathbf{z}_i - \mathbf{x}_0\| < \|\mathbf{h}\| \qquad \text{for all } 1 \leq i \leq n.$$

Therefore,

$$\frac{f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) - \langle \nabla f(\mathbf{x}_0), \mathbf{h} \rangle}{\|\mathbf{h}\|} = \sum_{i=1}^{n} \frac{h_i}{\|\mathbf{h}\|} \left(\frac{\partial f}{\partial x_i}(\mathbf{z}_i) - \frac{\partial f}{\partial x_i}(\mathbf{x}_0) \right).$$

Fixed $\varepsilon > 0$. For $1 \leq i \leq n$, since $f_{x_i} : B(\mathbf{x}_0, r) \to \mathbb{R}$ is continuous at $\mathbf{x}_0$, there exists $0 < \delta_i \leq r$ such that if $0 < \|\mathbf{z} - \mathbf{x}_0\| < \delta_i$, then

$$|f_{x_i}(\mathbf{z}) - f_{x_i}(\mathbf{x}_0)| < \frac{\varepsilon}{n}.$$

Take $\delta = \min\{\delta_1, \ldots, \delta_n\}$. Then $\delta > 0$. If $\|\mathbf{h}\| < \delta$, then for $1 \leq i \leq n$, $\|\mathbf{z}_i - \mathbf{x}_0\| < \|\mathbf{h}\| < \delta \leq \delta_i$. Thus,

$$|f_{x_i}(\mathbf{z_i}) - f_{x_i}(\mathbf{x}_0)| < \frac{\varepsilon}{n}.$$

This implies that

$$\left| \frac{f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) - \langle \nabla f(\mathbf{x}_0), \mathbf{h} \rangle}{\|\mathbf{h}\|} \right| \leq \sum_{i=1}^{n} \frac{|h_i|}{\|\mathbf{h}\|} \left| \frac{\partial f}{\partial x_i}(\mathbf{z}_i) - \frac{\partial f}{\partial x_i}(\mathbf{x}_0) \right|$$
$$< \varepsilon.$$

Hence,

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) - \langle \nabla f(\mathbf{x}_0), \mathbf{h} \rangle}{\|\mathbf{h}\|} = 0.$$

This proves that f is differentiable at $\mathbf{x}_0$.

Theorem 4.9 says that a function which has continuous partial derivatives is differentiable. This prompts us to make the following definition.

Definition 4.10 Continuously Differentiable

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ be a function defined on $\mathcal{O}$. We say that $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is continuously differentiable, or C^1, provided that it has partial derivatives that are continuous.

Theorem 4.9 says that a continuously differentiable function is differentiable. Analogously, we define C^k for any $k \geq 1$.

Definition 4.11 C^k Functions

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ be a function defined on $\mathcal{O}$. We say that $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is k-times continuously differentiable, or C^k, provided that it has all partial derivatives of order k, and each of them is continuous.

Definition 4.12 C^∞ Functions

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ be a function defined on $\mathcal{O}$. We say that $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is infinitely differentiable, or C^∞, provided that it is C^k for all positive integers k.

Proposition 4.10

Polynomials and rational functions are infinitely differentiable functions.

Sketch of Proof

A partial derivative of a rational function is still a rational function, which is continuous.

Obviously, for any $k \in \mathbb{Z}^+$, a C^{k+1} function is C^k.

Remark 4.8 Higher Order Differentiability

We can define second order differentiability in the following way. We say that a function $\mathbf{F} : \mathcal{O} \to \mathbb{R}$ is twice differentiable at a point $\mathbf{x}_0$ in $\mathcal{O}$ if there is a neighbourhood of $\mathbf{x}_0$ which $\mathbf{F}$ has first order partial derivatives, and each of them is differentiable at the point $\mathbf{x}_0$. Theorem 4.9 says that a C^2 function is twice differentiable.
Similarly, we can define higher order differentiability.

4.2.2 First Order Approximations

First we extend the concept of order of approximation to multivariable functions.

Definition 4.13 Order of Approximation

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let k be a positive integer. We say that the two functions $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ and $\mathbf{G} : \mathcal{O} \to \mathbb{R}^m$ are k^{th}-order of approximations of each other at $\mathbf{x}_0$ provided that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\mathbf{F}(\mathbf{x}_0+\mathbf{h}) - \mathbf{G}(\mathbf{x}_0+\mathbf{h})}{\|\mathbf{h}\|^k} = \mathbf{0}.$$

Recall that a mapping $\mathbf{G} : \mathcal{O} \to \mathbb{R}^m$ is a polynomial mapping of degree at most one if it has the form

$$\mathbf{G}(\mathbf{x}) = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n + b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n + b_2 \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n + b_m \end{bmatrix} = A\mathbf{x} + \mathbf{b},$$

where $A = [a_{ij}]$ and $\mathbf{b} = (b_1, \ldots, b_m)$. The mapping $\mathbf{G}$ is a linear transformation if and only if $\mathbf{b} = \mathbf{0}$.

The following theorem shows that first order approximation is closely related to differentiability. It is a consequence of Theorem 4.6.

Theorem 4.11 First Order Approximation Theorem

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ be a function defined on $\mathcal{O}$.

(a) If $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is continuous at $\mathbf{x}_0$, and there is a polynomial mapping $\mathbf{G} : \mathcal{O} \to \mathbb{R}^m$ of degree at most one which is a first order approximation of $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ at the point $\mathbf{x}_0$, then $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is differentiable at $\mathbf{x}_0$.

(b) If $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is differentiable at $\mathbf{x}_0$, then there is a unique polynomial mapping $\mathbf{G} : \mathcal{O} \to \mathbb{R}^m$ of degree at most one which is a first order approximation of $\mathbf{F}$ at $\mathbf{x}_0$. It is given by

$$\mathbf{G}(\mathbf{x}) = \mathbf{F}(\mathbf{x}_0) + \mathbf{DF}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0).$$

Proof

First we prove (a). Assume that $\mathbf{G} : \mathcal{O} \to \mathbb{R}^m$ is a polynomial mapping of degree at most one which is a first order approximation of $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ at the point $\mathbf{x}_0$. There exists an $m \times n$ matrix A and a vector $\mathbf{b}$ in $\mathbb{R}^m$ such that

$$\mathbf{G}(\mathbf{x}) = A\mathbf{x} + \mathbf{b}.$$

By assumption,

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\mathbf{F}(\mathbf{x}_0 + \mathbf{h}) - A(\mathbf{x}_0 + \mathbf{h}) - \mathbf{b}}{\|\mathbf{h}\|} = \mathbf{0}. \tag{4.7}$$

This implies that

$$\lim_{\mathbf{h}\to\mathbf{0}} (\mathbf{F}(\mathbf{x}_0 + \mathbf{h}) - A(\mathbf{x}_0 + \mathbf{h}) - \mathbf{b}) = \mathbf{0},$$

which gives

$$A\mathbf{x}_0 + \mathbf{b} = \lim_{\mathbf{h}\to\mathbf{0}} \mathbf{F}(\mathbf{x}_0 + \mathbf{h}) = \mathbf{F}(\mathbf{x}_0).$$

Substitute back into (4.7), we find that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\mathbf{F}(\mathbf{x}_0 + \mathbf{h}) - \mathbf{F}(\mathbf{x}_0) - A\mathbf{h}}{\|\mathbf{h}\|} = \mathbf{0}.$$

Since $\mathbf{T}(\mathbf{h}) = A\mathbf{h}$ is a linear transformation, this shows that $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is differentiable at $\mathbf{x}_0$.

Next, we prove (b). If $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is differentiable at $\mathbf{x}_0$, Theorem 4.6 says that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\mathbf{F}(\mathbf{x}_0 + \mathbf{h}) - \mathbf{F}(\mathbf{x}_0) - \mathbf{DF}(\mathbf{x}_0)\mathbf{h}}{\|\mathbf{h}\|} = \mathbf{0}.$$

This precisely means that the polynomial mapping $\mathbf{G} : \mathcal{O} \to \mathbb{R}^m$,

$$\mathbf{G}(\mathbf{x}) = \mathbf{F}(\mathbf{x}_0) + \mathbf{DF}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0),$$

is a first order approximation of $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ at $\mathbf{x}_0$. By definition, the polynomial mapping $\mathbf{G}$ has degree at most one. The uniqueness of $\mathbf{G}$ is also asserted in Theorem 4.6.

Remark 4.9

The first order approximation theorem says that if the function $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is differentiable at the point $\mathbf{u}$, then there is a unique polynomial mapping $\mathbf{G} : \mathcal{O} \to \mathbb{R}^m$ of degree at most one which is a first order approximation of $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ at the point $\mathbf{u}$. The components of the mapping $\mathbf{G} : \mathcal{O} \to \mathbb{R}^m$ are given by

$$G_j(x_1, \ldots, x_n) = F_j(u_1, \ldots, u_n) + \sum_{i=1}^{n} \frac{\partial F_j}{\partial x_i}(u_1, \ldots, u_n)(x_i - u_i).$$

Notice that this is a (generalization) of Taylor polynomial of order 1.

Example 4.14

Let $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^2$ be the function defined as

$$\mathbf{F}(x, y, z) = (xyz^2, x + 2y + 3z),$$

and let $\mathbf{x}_0 = (1, -1, 1)$. Find a vector $\mathbf{b}$ in $\mathbb{R}^2$ and a 2×3 matrix A such that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\mathbf{F}(\mathbf{x}_0 + \mathbf{h}) - A\mathbf{h} - \mathbf{b}}{\|\mathbf{h}\|} = \mathbf{0}.$$

Solution

The function $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^2$ is itself a polynomial mapping. Hence, it is differentiable. The derivative matrix is given by

$$\mathbf{DF}(\mathbf{x}) = \begin{bmatrix} yz^2 & xz^2 & 2xyz \\ 1 & 2 & 3 \end{bmatrix}.$$

By the first order approximation theorem, $\mathbf{b} = \mathbf{F}(\mathbf{x}_0) = (-1, 2)$ and

$$A = \mathbf{DF}(1, -1, 1) = \begin{bmatrix} -1 & 1 & -2 \\ 1 & 2 & 3 \end{bmatrix}.$$

Example 4.15

Determine whether the limit $\lim\limits_{(x,y)\to(0,0)} \dfrac{e^{x+2y} - 1 - x - 2y}{\sqrt{x^2+y^2}}$ exists.

Solution

Let $f(x, y) = e^{x+2y}$. Then

$$\frac{\partial f}{\partial x}(x, y) = e^{x+2y}, \qquad \frac{\partial f}{\partial y}(x, y) = 2e^{x+2y}.$$

It follows that

$$f(0, 0) = 1, \quad \frac{\partial f}{\partial x}(0, 0) = 1, \quad \frac{\partial f}{\partial y}(0, 0) = 2.$$

Since the function $g(x, y) = x + 2y$ is continuous and the exponential function is also continuous, f has continuous first order partial derivatives. Hence, f is differentiable. By first order approximation theorem,

$$\lim_{(x,y)\to(0,0)} \frac{f(x, y) - f(0, 0) - x\dfrac{\partial f}{\partial x}(0, 0) - y\dfrac{\partial f}{\partial y}(0, 0)}{\sqrt{x^2+y^2}} = 0.$$

Since

$$f(x,y) - f(0,0) - x\frac{\partial f}{\partial x}(0,0) - y\frac{\partial f}{\partial y}(0,0) = e^{x+2y} - 1 - x - 2y,$$

we find that

$$\lim_{(x,y)\to(0,0)} \frac{e^{x+2y} - 1 - x - 2y}{\sqrt{x^2+y^2}} = 0.$$

4.2.3 Tangent Planes

The tangent plane to a graph is closely related to the concept of differentiability and first order approximations. Recall that the graph of a function $f : \mathcal{O} \to \mathbb{R}$ defined on a subset of $\mathbb{R}^n$ is the subset of $\mathbb{R}^{n+1}$ consists of all the points of the form $(\mathbf{x}, f(\mathbf{x}))$ where $\mathbf{x} \in \mathcal{O}$.

Definition 4.14 Tangent Planes

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $f : \mathcal{O} \to \mathbb{R}$ be a function defined on $\mathcal{O}$. The graph of f has a tangent plane at $\mathbf{x}_0$ if it is differentiable at $\mathbf{x}_0$. In this case, the tangent plane is the hyperplane of $\mathbb{R}^{n+1}$ that satisfies the equation

$$x_{n+1} = f(\mathbf{x}_0) + \langle \nabla f(\mathbf{x}_0), \mathbf{x} - \mathbf{x}_0 \rangle, \qquad \text{where } \mathbf{x} = (x_1, \ldots, x_n).$$

The tangent plane is the graph of the polynomial function of degree at most one which is the first order approximation of the function f at the point $\mathbf{x}_0$.

Example 4.16

Find the equation of the tangent plane to the graph of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x,y) = x^2 + 4xy + 5y^2$ at the point where $(x,y) = (1,-1)$.

Solution

The function f is a polynomial. Hence, it is a differentiable function with

$$\nabla f(x,y) = (2x + 4y, 4x + 10y).$$

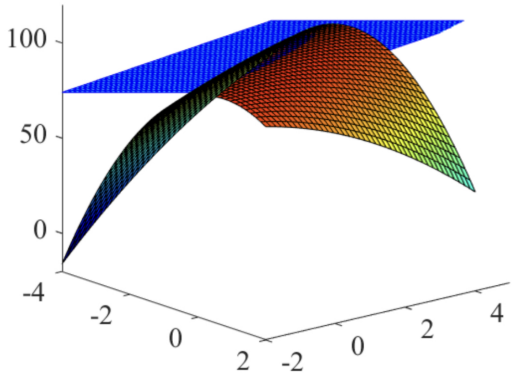


Figure 4.7: The tangent plane to the graph of a function.

From this, we find that $\nabla f(1,-1) = (-2,-6)$. Together with $f(1,-1) = 2$, we find that the equation of the tangent plane to the graph of f at the point where $(x,y) = (1,-1)$ is

$$z = 2 - 2(x-1) - 6(y+1) = -2x - 6y - 2.$$

4.2.4 Directional Derivatives

As we mentioned before, the partial derivatives measure the rate of change of the function when it varies along the directions of the coordinate axes. To capture the rate of change of a function along other directions, we define the concept of directional derivatives. Notice that a direction in $\mathbb{R}^n$ is specified by a *unit* vector.

Definition 4.15 Directional Derivatives

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ be a function defined on $\mathcal{O}$. Given a **unit** vector $\mathbf{u}$ in $\mathbb{R}^n$, we say that $\mathbf{F}$ has directional derivative in the direction of $\mathbf{u}$ at the point $\mathbf{x}_0$ provided that the limit

$$\lim_{h\to 0} \frac{\mathbf{F}(\mathbf{x}_0 + h\mathbf{u}) - \mathbf{F}(\mathbf{x}_0)}{h}$$

exists. This limit, denoted as $\mathbf{D}_{\mathbf{u}}\mathbf{F}(\mathbf{x}_0)$, is called the directional derivative of $\mathbf{F}$ in the direction of $\mathbf{u}$ at the point $\mathbf{x}_0$.

When $m = 1$, it is customary to denote the directional derivative of $f : \mathcal{O} \to \mathbb{R}$ in the direction of $\mathbf{u}$ at the point $\mathbf{x}_0$ as $D_{\mathbf{u}} f(\mathbf{x}_0)$.

Remark 4.10

For any nonzero vector $\mathbf{v}$ in $\mathbb{R}^n$, we can also define $\mathbf{D}_{\mathbf{v}}\mathbf{F}(\mathbf{x}_0)$ as

$$\mathbf{D}_{\mathbf{v}}\mathbf{F}(\mathbf{x}_0) = \lim_{h \to 0} \frac{\mathbf{F}(\mathbf{x}_0 + h\mathbf{v}) - \mathbf{F}(\mathbf{x}_0)}{h}.$$

However, we will not call it a directional derivative unless $\mathbf{v}$ is a unit vector.

Remark 4.11

From the definition, it is obvious that when $\mathbf{u}$ is one of the standard unit vectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$, then the directional derivative in the direction of $\mathbf{u}$ is a partial derivative. More precisely,

$$\mathbf{D}_{\mathbf{e}_i}\mathbf{F}(\mathbf{x}_0) = \frac{\partial \mathbf{F}}{\partial x_i}(\mathbf{x}_0), \qquad 1 \leq i \leq n.$$

The following is obvious.

Proposition 4.12

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ be a function defined on $\mathcal{O}$. Given a nonzero vector $\mathbf{v}$ in $\mathbb{R}^n$, $\mathbf{D}_{\mathbf{v}}\mathbf{F}(\mathbf{x}_0)$ exists if only if $D_{\mathbf{v}} F_j(\mathbf{x}_0)$ exists for all $1 \leq j \leq m$. Moreover,

$$\mathbf{D}_{\mathbf{v}}\mathbf{F}(\mathbf{x}_0) = \left(D_{\mathbf{v}} F_1(\mathbf{x}_0), D_{\mathbf{v}} F_2(\mathbf{x}_0), \ldots, D_{\mathbf{v}} F_m(\mathbf{x}_0)\right).$$

Example 4.17

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function defined as

$$f(x, y) = x^2 y.$$

Given that $\mathbf{v} = (v_1, v_2)$ is a nonzero vector in $\mathbb{R}^2$, find $D_{\mathbf{v}} f(3, 2)$.

Solution

By definition,

$$D_{\mathbf{v}}f(3,2) = \lim_{h\to 0}\frac{f(3+hv_1, 2+hv_2) - f(3,2)}{h} = g'(0),$$

where

$$g(h) = f(3+hv_1, 2+hv_2) = (3+hv_1)^2(2+hv_2).$$

Since

$$g'(h) = 2v_1(3+hv_1)(2+hv_2) + v_2(3+hv_1)^2,$$

we find that

$$D_{\mathbf{v}}f(3,2) = g'(0) = 12v_1 + 9v_2.$$

Take $\mathbf{v} = \mathbf{e}_1 = (1,0)$ and $\mathbf{v} = \mathbf{e}_2 = (0,1)$ respectively, we find that $f_x(3,2) = 12$ and $f_y(3,2) = 9$. For general $\mathbf{v} = (v_1, v_2)$, we notice that

$$D_{\mathbf{v}}f(3,2) = \langle \nabla f(3,2), \mathbf{v}\rangle.$$

Example 4.18

Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined as

$$f(x,y) = \begin{cases} \dfrac{xy}{x^2+y^2}, & \text{if } (x,y) \neq (0,0), \\ 0, & \text{if } (x,y) = (0,0) \end{cases}$$

in Example 4.6. Find all the nonzero vectors $\mathbf{v}$ for which $D_{\mathbf{v}}f(0,0)$ exists.

Solution

Given a nonzero vector $\mathbf{v} = (v_1, v_2)$, $v_1^2 + v_2^2 \neq 0$. By definition,

$$D_{\mathbf{v}}f(0,0) = \lim_{h\to 0}\frac{f(hv_1, hv_2) - f(0,0)}{h} = \lim_{h\to 0}\frac{1}{h}\frac{v_1v_2}{v_1^2+v_2^2}.$$

This limit exists if and only if $v_1v_2 = 0$, which is the case if $v_1 = 0$ or $v_2 = 0$.

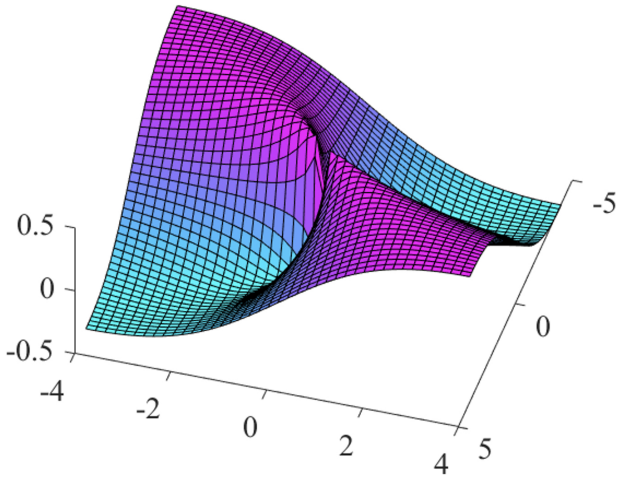


Figure 4.8: The function $f(x, y)$ in Example 4.18.

Example 4.19

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function defined as

$$f(x,y) = \begin{cases} \dfrac{y\sqrt{x^2+y^2}}{|x|}, & \text{if } x \neq 0, \\ 0, & \text{if } x = 0. \end{cases}$$

Find all the nonzero vectors $\mathbf{v}$ for which $D_{\mathbf{v}} f(0,0)$ exists.

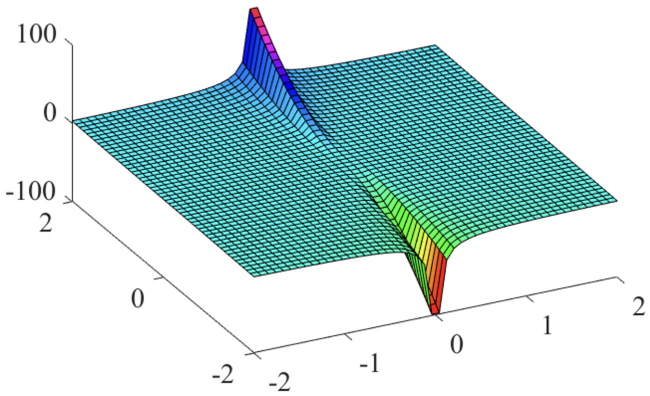


Figure 4.9: The function $f(x, y)$ in Example 4.19.

Solution

Given a nonzero vector $\mathbf{v} = (v_1, v_2)$, we consider two cases.

Case I: $v_1 = 0$.

Then $\mathbf{v} = (0, v_2)$. In this case,

$$D_{\mathbf{v}} f(0,0) = \lim_{h \to 0} \frac{f(0, hv_2) - f(0,0)}{h} = \lim_{h \to 0} \frac{0 - 0}{h} = 0.$$

Case 2: $v_1 \neq 0$.

$$\begin{aligned} D_{\mathbf{v}} f(0,0) &= \lim_{h \to 0} \frac{f(hv_1, hv_2) - f(0,0)}{h} \\ &= \lim_{h \to 0} \frac{1}{h} \frac{hv_2}{|hv_1|} \sqrt{h^2(v_1^2 + v_2^2)} \\ &= \frac{v_2 \sqrt{v_1^2 + v_2^2}}{|v_1|}. \end{aligned}$$

We conclude that $D_{\mathbf{v}} f(0,0)$ exists for all nonzero vectors $\mathbf{v}$.

Remark 4.12

For the function considered in Example 4.19, by taking $\mathbf{v}$ to be $(1,0)$ and $(0,1)$ respectively, we find that $f_x(0,0) = 0$ and $f_y(0,0) = 0$. Notice that

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{f(\mathbf{h}) - f(\mathbf{0}) - \langle \nabla f(\mathbf{0}), \mathbf{h} \rangle}{\|\mathbf{h}\|} = \lim_{\mathbf{h} \to \mathbf{0}} \frac{h_2}{|h_1|}.$$

This limit does not exist. By Corollary 4.7, f is not differentiable at $(0,0)$. This gives an example of a function which is not differentiable at $(0,0)$ but has directional derivatives at $(0,0)$ in all directions. In fact, one can show that f is not continuous at $(0,0)$.

The following theorem says that differentiability of a function implies existence of directional derivatives.

Theorem 4.13

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ be a function defined on $\mathcal{O}$. If $\mathbf{F}$ is differentiable at $\mathbf{x}_0$, then for any nonzero vector $\mathbf{v}$, $\mathbf{D}_\mathbf{v}\mathbf{F}(\mathbf{x}_0)$ exists and

$$\mathbf{D}_\mathbf{v}\mathbf{F}(\mathbf{x}_0) = \mathbf{DF}(\mathbf{x}_0)\mathbf{v} = \begin{bmatrix} \langle \nabla F_1(\mathbf{x}_0), \mathbf{v} \rangle \\ \langle \nabla F_2(\mathbf{x}_0), \mathbf{v} \rangle \\ \vdots \\ \langle \nabla F_m(\mathbf{x}_0), \mathbf{v} \rangle \end{bmatrix}.$$

Proof

Again, it is sufficient to consider a function $f : \mathcal{O} \to \mathbb{R}$ with codomain $\mathbb{R}$. By definition, $D_\mathbf{v} f(\mathbf{x}_0)$ is given by the limit

$$\lim_{h \to 0} \frac{f(\mathbf{x}_0 + h\mathbf{v}) - f(\mathbf{x}_0)}{h}$$

if it exists. Since f is differentiable at $\mathbf{x}_0$, it has partial derivatives at $\mathbf{x}_0$ and

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) - \langle \nabla f(\mathbf{x}_0), \mathbf{h} \rangle}{\|\mathbf{h}\|} = 0.$$

As $h \to 0$, $h\mathbf{v} \to \mathbf{0}$. By limit law for composite functions, we find that

$$\lim_{h \to 0} \frac{f(\mathbf{x}_0 + h\mathbf{v}) - f(\mathbf{x}_0) - \langle \nabla f(\mathbf{x}_0), h\mathbf{v} \rangle}{|h|\|\mathbf{v}\|} = 0.$$

This implies that

$$\lim_{h \to 0} \frac{f(\mathbf{x}_0 + h\mathbf{v}) - f(\mathbf{x}_0) - h\langle \nabla f(\mathbf{x}_0), \mathbf{v} \rangle}{h} = 0.$$

Thus,

$$D_\mathbf{v} f(\mathbf{x}_0) = \lim_{h \to 0} \frac{f(\mathbf{x}_0 + h\mathbf{v}) - f(\mathbf{x}_0)}{h} = \langle \nabla f(\mathbf{x}_0), \mathbf{v} \rangle.$$

Example 4.20

Consider the function $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$ defined as $\mathbf{F}(x, y) = (x^2y, xy^2)$. Find $\mathbf{D_v F}(2, 3)$ when $\mathbf{v} = (-1, 2)$.

Solution

Since $\mathbf{F}$ is a polynomial mapping, it is differentiable. The derivative matrix is $\mathbf{DF}(x, y) = \begin{bmatrix} 2xy & x^2 \\ y^2 & 2xy \end{bmatrix}$. Therefore,

$$\mathbf{D_v F}(2, 3) = \mathbf{DF}(2, 3) \begin{bmatrix} -1 \\ 2 \end{bmatrix} = \begin{bmatrix} 12 & 4 \\ 9 & 12 \end{bmatrix} \begin{bmatrix} -1 \\ 2 \end{bmatrix} = \begin{bmatrix} -4 \\ 15 \end{bmatrix}.$$

Theorem 4.13 can be used to determine the direction which a differentiable function increase fastest at a point.

Corollary 4.14

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $f : \mathcal{O} \to \mathbb{R}$ be a function defined on $\mathcal{O}$. If f is differentiable at $\mathbf{x}_0$ and $\nabla f(\mathbf{x}_0) \neq \mathbf{0}$, then at the point $\mathbf{x}_0$, the function f increases fastest in the direction of $\nabla f(\mathbf{x}_0)$.

Proof

Let $\mathbf{u}$ be a unit vector. Then the rate of change of the function f at the point $\mathbf{x}_0$ in the direction of $\mathbf{u}$ is given by

$$D_{\mathbf{u}} f(\mathbf{x}_0) = \langle \nabla f(\mathbf{x}_0), \mathbf{u} \rangle.$$

By Cauchy-Schwarz inequality,

$$\langle \nabla f(\mathbf{x}_0), \mathbf{u} \rangle \leq \|\nabla f(\mathbf{x}_0)\| \|\mathbf{u}\| = \|\nabla f(\mathbf{x}_0)\|,$$

and the equality holds if and only if $\mathbf{u}$ has the same direction as $\nabla f(\mathbf{x}_0)$.

Exercises 4.2

Question 1

Let $f : \mathbb{R}^3 \to \mathbb{R}$ be the function defined as

$$f(x, y, z) = xe^{y^2+4z}.$$

Find a vector $\mathbf{c}$ in $\mathbb{R}^3$ and a constant b such that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{f(\mathbf{x}_0 + \mathbf{h}) - \langle \mathbf{c}, \mathbf{h}\rangle - b}{\|\mathbf{h}\|} = 0,$$

where $\mathbf{x}_0 = (3, 2, -1)$.

Question 2

Let $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^3$ be the function defined as

$$\mathbf{F}(x, y) = (x^2 + 4y^2, 7xy, 2x + y).$$

Find a polynomial mapping $\mathbf{G} : \mathbb{R}^2 \to \mathbb{R}^3$ of degree at most one which is a first order approximation of $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^3$ at the point $(1, -1)$.

Question 3

Let $\mathbf{x}_0 = (1, 2, 0, -1)$, and let $\mathbf{F} : \mathbb{R}^4 \to \mathbb{R}^3$ be the function defined as

$$\mathbf{F}(x_1, x_2, x_3, x_4) = \left(x_2x_3^2,\ x_3x_4^3 + x_2,\ x_4 + 2x_1 + 1\right).$$

Find a 3×4 matrix A and a vector $\mathbf{b}$ in $\mathbb{R}^3$ such that

$$\lim_{\mathbf{x}\to\mathbf{x}_0} \frac{\mathbf{F}(\mathbf{x}) - A\mathbf{x} - \mathbf{b}}{\|\mathbf{x} - \mathbf{x}_0\|} = \mathbf{0}.$$

Question 4

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function defined as

$$f(x, y) = \sin(x^2 + y) + 5xy^2.$$

Find $D_{\mathbf{v}} f(1, -1)$ for any nonzero vector $\mathbf{v} = (v_1, v_2)$.

Question 5

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function defined as

$$f(x, y) = \begin{cases} \dfrac{x^2 y^2}{x^2 + y^2}, & \text{if } (x, y) \neq (0, 0) \\ 0, & \text{if } (x, y) = (0, 0). \end{cases}$$

Show that $f : \mathbb{R}^2 \to \mathbb{R}$ is continuously differentiable.

Question 6

Find the equation of the tangent plane to the graph of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = 4x^2 + 3xy - y^2$ at the point where $(x, y) = (2, -1)$.

Question 7

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function defined as

$$f(x, y) = \begin{cases} \dfrac{x^2 y}{x^2 + y^2}, & \text{if } (x, y) \neq (0, 0), \\ 0, & \text{if } (x, y) = (0, 0). \end{cases}$$

(a) Show that $f : \mathbb{R}^2 \to \mathbb{R}$ is continuous.

(b) Show that $f : \mathbb{R}^2 \to \mathbb{R}$ has partial derivatives.

(c) Show that $f : \mathbb{R}^2 \setminus \{(0, 0)\} \to \mathbb{R}$ is differentiable.

(d) Show that $f : \mathbb{R}^2 \to \mathbb{R}$ is not differentiable at $(0, 0)$.

(e) Find all the nonzero vectors $\mathbf{v} = (v_1, v_2)$ for which $D_{\mathbf{v}} f(0, 0)$ exists.

Question 8

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function defined as

$$f(x,y) = \begin{cases} \dfrac{|x|\sqrt{x^2+y^2}}{y}, & \text{if } y \neq 0, \\ 0, & \text{if } y = 0. \end{cases}$$

(a) Show that $f : \mathbb{R}^2 \to \mathbb{R}$ is not continuous at $(0,0)$.

(b) Show that $D_{\mathbf{v}} f(0,0)$ exists for all nonzero vectors $\mathbf{v}$.

Question 9

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function defined as

$$f(x,y) = \begin{cases} (x^2+y^2)\sin\left(\dfrac{1}{\sqrt{x^2+y^2}}\right), & \text{if } (x,y) \neq (0,0), \\ 0, & \text{if } (x,y) = (0,0). \end{cases}$$

(a) Show that $f : \mathbb{R}^2 \to \mathbb{R}$ is differentiable at $(0,0)$.

(b) Show that $f : \mathbb{R}^2 \to \mathbb{R}$ is not continuously differentiable at $(0,0)$.

4.3 The Chain Rule and the Mean Value Theorem

In volume I, we have seen that the chain rule plays an important role in calculating the derivative of a composite function. Given that $f : (a, b) \to \mathbb{R}$ and $g : (c, d) \to \mathbb{R}$ are functions such that $f((a, b)) \subset (c, d)$, the chain rule says that if f is differentiable at x_0, g is differentiable at $y_0 = f(x_0)$, then the composite function $(g \circ f) : (a, b) \to \mathbb{R}$ is differentiable at x_0, and

$$(g \circ f)'(x_0) = g'(f(x_0))f'(x_0).$$

For multivariable functions, the chain rule takes the following form.

Theorem 4.15 The Chain Rule

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $\mathcal{U}$ be an open subset of $\mathbb{R}^k$. Assume that $\mathbf{F} : \mathcal{O} \to \mathbb{R}^k$ and $\mathbf{G} : \mathcal{U} \to \mathbb{R}^m$ are functions such that $\mathbf{F}(\mathcal{O}) \subset \mathcal{U}$. If $\mathbf{F}$ is differentiable at $\mathbf{x}_0$, $\mathbf{G}$ is differentiable at $\mathbf{y}_0 = \mathbf{F}(\mathbf{x}_0)$, then the composite function $\mathbf{H} = (\mathbf{G} \circ \mathbf{F}) : \mathcal{O} \to \mathbb{R}^m$ is differentiable at $\mathbf{x}_0$ and

$$\mathbf{DH}(\mathbf{x}_0) = \mathbf{D}(\mathbf{G} \circ \mathbf{F})(\mathbf{x}_0) = \mathbf{DG}(\mathbf{F}(\mathbf{x}_0))\mathbf{DF}(\mathbf{x}_0).$$

Notice that on the right hand side, $\mathbf{DG}(\mathbf{F}(\mathbf{x}_0))$ is an $m \times k$ matrix, $\mathbf{DF}(\mathbf{x}_0)$ is an $k \times n$ matrix. Hence, the product $\mathbf{DG}(\mathbf{F}(\mathbf{x}_0))\mathbf{DF}(\mathbf{x}_0)$ makes sense, and it is an $m \times n$ matrix, which is the correct size for the derivative matrix $\mathbf{DH}(\mathbf{x}_0)$.

Let us spell out more explicitly. Assume that

$$\begin{aligned}
&\mathbf{F}(x_1, x_2, \ldots, x_n) \\
&= (F_1(x_1, x_2, \ldots, x_n), F_2(x_1, x_2, \ldots, x_n), \ldots, F_k(x_1, x_2, \ldots, x_n)), \\
&\mathbf{G}(y_1, y_2, \ldots, y_k) \\
&= (G_1(y_1, y_2, \ldots, y_k), G_2(y_1, y_2, \ldots, y_k), \ldots, G_m(y_1, y_2, \ldots, y_k)), \\
&\mathbf{H}(x_1, x_2, \ldots, x_n) \\
&= (H_1(x_1, x_2, \ldots, x_n), H_2(x_1, x_2, \ldots, x_n), \ldots, H_m(x_1, x_2, \ldots, x_n)).
\end{aligned}$$

Then for $1 \leq j \leq m$,

$$\begin{aligned}
&H_j(x_1, x_2, \ldots, x_n) \\
&= G_j\left(F_1(x_1, x_2, \ldots, x_n), F_2(x_1, x_2, \ldots, x_n), \ldots, F_k(x_1, x_2, \ldots, x_n)\right).
\end{aligned}$$

For $1 \leq l \leq k$, let

$$y_l = F_l\left(x_1, x_2, \ldots, x_n\right).$$

The chain rule says that if $1 \leq q \leq n$,

$$\begin{aligned}\frac{\partial H_j}{\partial x_q}(x_1, x_2, \ldots, x_n) &= \sum_{l=1}^{k} \frac{\partial G_j}{\partial y_l}(y_1, y_2, \ldots, y_k)\frac{\partial F_l}{\partial x_q}(x_1, x_2, \ldots, x_n) \\ &= \frac{\partial G_j}{\partial y_1}(y_1, y_2, \ldots, y_k)\frac{\partial F_1}{\partial x_q}(x_1, x_2, \ldots, x_n) \\ &\quad + \frac{\partial G_j}{\partial y_2}(y_1, y_2, \ldots, y_k)\frac{\partial F_2}{\partial x_q}(x_1, x_2, \ldots, x_n) \\ &\qquad\qquad \vdots \\ &\quad + \frac{\partial G_j}{\partial y_k}(y_1, y_2, \ldots, y_k)\frac{\partial F_k}{\partial x_q}(x_1, x_2, \ldots, x_n).\end{aligned}$$

Namely, to differentiate $H_j = G_j \circ \mathbf{F}$ with respect to x_q, we differentiate G_j with respect to each of the variables $y_1, \ldots, y_k$, multiply each by the partial derivatives of $F_1, \ldots, F_k$ with respect to x_q, then take the sum.

Let us illustrate this with a simple example.

Example 4.21

Consider the function $h : \mathbb{R}^2 \to \mathbb{R}$ defined as

$$h(x, y) = \sin(2x + 3y) + e^{xy}.$$

It is straightforward to find that

$$\frac{\partial h}{\partial x} = 2\cos(2x + 3y) + ye^{xy}, \qquad \frac{\partial h}{\partial y} = 3\cos(2x + 3y) + xe^{xy}.$$

Notice that we can write $h = g \circ \mathbf{F}$, where $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$ is the function

$$\mathbf{F}(x, y) = (2x + 3y, xy),$$

and $g : \mathbb{R}^2 \to \mathbb{R}$ is the function

$$g(u, v) = \sin u + e^v.$$

Obviously, $\mathbf{F}$ and g are continuously differentiable functions.

$$\mathbf{DF}(x,y) = \begin{bmatrix} 2 & 3 \\ y & x \end{bmatrix}, \qquad Dg(u,v) = \begin{bmatrix} \cos u & e^v \end{bmatrix}.$$

Taking $u = 2x + 3y$ and $v = xy$, we find that

$$\begin{aligned} Dg(u,v)\mathbf{DF}(x,y) &= \begin{bmatrix} \cos(2x+3y) & e^{xy} \end{bmatrix} \begin{bmatrix} 2 & 3 \\ y & x \end{bmatrix} \\ &= \begin{bmatrix} 2\cos(2x+3y) + ye^{xy} & 3\cos(2x+3y) + xe^{xy} \end{bmatrix} \\ &= Dh(x,y). \end{aligned}$$

Now let us prove the chain rule.

Proof of the Chain Rule

Since $\mathbf{F}$ is differentiable at $\mathbf{x}_0$ and $\mathbf{G}$ is differentiable at $\mathbf{y}_0 = \mathbf{F}(\mathbf{x}_0)$, $\mathbf{DF}(\mathbf{x}_0)$ and $\mathbf{DG}(\mathbf{y}_0)$ exist. There exists positive numbers r_1 and r_2 such that $B(\mathbf{x}_0, r_1) \subset \mathcal{O}$ and $B(\mathbf{y}_0, r_2) \subset \mathcal{U}$. Let

$$\begin{aligned} \boldsymbol{\varepsilon}_1(\mathbf{h}) &= \frac{\mathbf{F}(\mathbf{x}_0 + \mathbf{h}) - \mathbf{F}(\mathbf{x}_0) - \mathbf{DF}(\mathbf{x}_0)\mathbf{h}}{\|\mathbf{h}\|}, \qquad \mathbf{h} \in B(\mathbf{0}, r_1), \\ \boldsymbol{\varepsilon}_2(\mathbf{v}) &= \frac{\mathbf{G}(\mathbf{y}_0 + \mathbf{v}) - \mathbf{G}(\mathbf{y}_0) - \mathbf{DG}(\mathbf{y}_0)\mathbf{v}}{\|\mathbf{v}\|}, \qquad \mathbf{v} \in B(\mathbf{0}, r_2). \end{aligned}$$

Since $\mathbf{F}$ is differentiable at $\mathbf{x}_0$ and $\mathbf{G}$ is differentiable at $\mathbf{y}_0$,

$$\lim_{\mathbf{h}\to\mathbf{0}} \boldsymbol{\varepsilon}_1(\mathbf{h}) = \mathbf{0}, \qquad \lim_{\mathbf{v}\to\mathbf{0}} \boldsymbol{\varepsilon}_2(\mathbf{v}) = \mathbf{0}.$$

There exist positive constants c_1 and c_2 such that

$$\|\mathbf{DF}(\mathbf{x}_0)\mathbf{h}\| \le c_1\|\mathbf{h}\| \qquad \text{for all } \mathbf{h} \in \mathbb{R}^n,$$

$$\|\mathbf{DG}(\mathbf{y}_0)\mathbf{v}\| \le c_2\|\mathbf{v}\| \qquad \text{for all } \mathbf{v} \in \mathbb{R}^k.$$

Now since $\mathbf{F}$ is differentiable at $\mathbf{x}_0$, it is continuous at $\mathbf{x}_0$. Hence, there exists a positive number r such that $r \le r_1$ and $\mathbf{F}(B(\mathbf{x}_0, r)) \subset B(\mathbf{y}_0, r_2)$.

For $\mathbf{h} \in B(\mathbf{0}, r)$, let

$$\mathbf{v} = \mathbf{F}(\mathbf{x}_0 + \mathbf{h}) - \mathbf{F}(\mathbf{x}_0).$$

Then $\mathbf{v} \in B(\mathbf{0}, r_2)$ and

$$\mathbf{v} = \mathbf{DF}(\mathbf{x}_0)\mathbf{h} + \|\mathbf{h}\|\boldsymbol{\varepsilon}_1(\mathbf{h}).$$

It follows that

$$\|\mathbf{v}\| \leq \|\mathbf{DF}(\mathbf{x}_0)\mathbf{h}\| + \|\mathbf{h}\|\|\boldsymbol{\varepsilon}_1(\mathbf{h})\| \leq \|\mathbf{h}\| \left(c_1 + \|\boldsymbol{\varepsilon}_1(\mathbf{h})\|\right).$$

In particular, we find that when $\mathbf{h} \to \mathbf{0}$, $\mathbf{v} \to \mathbf{0}$. Now,

$$\begin{aligned}
&\mathbf{H}(\mathbf{x}_0 + \mathbf{h}) - \mathbf{H}(\mathbf{x}_0) \\
&= \mathbf{G}(\mathbf{F}(\mathbf{x}_0 + \mathbf{h})) - \mathbf{G}(\mathbf{F}(\mathbf{x}_0)) \\
&= \mathbf{G}(\mathbf{y}_0 + \mathbf{v}) - \mathbf{G}(\mathbf{y}_0) \\
&= \mathbf{DG}(\mathbf{y}_0)\mathbf{v} + \|\mathbf{v}\|\boldsymbol{\varepsilon}_2(\mathbf{v}) \\
&= \mathbf{DG}(\mathbf{y}_0)\mathbf{DF}(\mathbf{x}_0)\mathbf{h} + \|\mathbf{h}\|\mathbf{DG}(\mathbf{y}_0)\boldsymbol{\varepsilon}_1(\mathbf{h}) + \|\mathbf{v}\|\boldsymbol{\varepsilon}_2(\mathbf{v}).
\end{aligned}$$

Therefore, for $\mathbf{h} \in B(\mathbf{0}, r) \setminus \{\mathbf{0}\}$,

$$\frac{\mathbf{H}(\mathbf{x}_0 + \mathbf{h}) - \mathbf{H}(\mathbf{x}_0) - \mathbf{DG}(\mathbf{y}_0)\mathbf{DF}(\mathbf{x}_0)\mathbf{h}}{\|\mathbf{h}\|} = \mathbf{DG}(\mathbf{y}_0)\boldsymbol{\varepsilon}_1(\mathbf{h}) + \frac{\|\mathbf{v}\|}{\|\mathbf{h}\|}\boldsymbol{\varepsilon}_2(\mathbf{v}).$$

This implies that

$$\begin{aligned}
&\left\| \frac{\mathbf{H}(\mathbf{x}_0 + \mathbf{h}) - \mathbf{H}(\mathbf{x}_0) - \mathbf{DG}(\mathbf{y}_0)\mathbf{DF}(\mathbf{x}_0)\mathbf{h}}{\|\mathbf{h}\|} \right\| \\
&\leq \|\mathbf{DG}(\mathbf{y}_0)\boldsymbol{\varepsilon}_1(\mathbf{h})\| + \frac{\|\mathbf{v}\|}{\|\mathbf{h}\|}\|\boldsymbol{\varepsilon}_2(\mathbf{v})\| \\
&\leq c_2\|\boldsymbol{\varepsilon}_1(\mathbf{h})\| + \left(c_1 + \|\boldsymbol{\varepsilon}_1(\mathbf{h})\|\right)\|\boldsymbol{\varepsilon}_2(\mathbf{v})\|.
\end{aligned}$$

Since $\mathbf{v} \to \mathbf{0}$ when $\mathbf{h} \to \mathbf{0}$, we find that $\boldsymbol{\varepsilon}_2(\mathbf{v}) \to \mathbf{0}$ when $\mathbf{h} \to \mathbf{0}$. Thus, we find that

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{\mathbf{H}(\mathbf{x}_0 + \mathbf{h}) - \mathbf{H}(\mathbf{x}_0) - \mathbf{DG}(\mathbf{y}_0)\mathbf{DF}(\mathbf{x}_0)\mathbf{h}}{\|\mathbf{h}\|} = \mathbf{0}.$$

This concludes that $\mathbf{H}$ is differentiable at $\mathbf{x}_0$ and

$$\mathbf{DH}(\mathbf{x}_0) = \mathbf{DG}(\mathbf{y}_0)\mathbf{DF}(\mathbf{x}_0).$$

Example 4.22

Let $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^2$ be the function defined as

$$\mathbf{F}(x, y, z) = (x^2 + 4y^2 + 9z^2, xyz).$$

Find a vector $\mathbf{b}$ in $\mathbb{R}^2$ and a 2×3 matrix A such that

$$\lim_{(u,v,w)\to(1,-1,0)} \frac{\mathbf{F}(2u + v, v + w, u + w) - \mathbf{b} - A\mathbf{p}}{\sqrt{(u-1)^2 + (v+1)^2 + w^2}} = \mathbf{0}, \quad \text{where } \mathbf{p} = \begin{bmatrix} u \\ v \\ w \end{bmatrix}.$$

Solution

Let $\mathbf{p}_0 = (1, -1, 0)$, and let $\mathbf{G} : \mathbb{R}^3 \to \mathbb{R}^3$ be the mapping

$$\mathbf{G}(u, v, w) = (2u + v, v + w, u + w).$$

Then $\mathbf{H}(\mathbf{p}) = \mathbf{H}(u, v, w) = \mathbf{F}(2u + v, v + w, u + w) = (\mathbf{F} \circ \mathbf{G})(u, v, w)$. Notice that $\mathbf{F}$ and $\mathbf{G}$ are polynomial mappings. Hence, they are infinitely differentiable. To have

$$\begin{aligned} &\lim_{\mathbf{p}\to\mathbf{p}_0} \frac{\mathbf{H}(\mathbf{v}) - \mathbf{b} - A\mathbf{p}}{\|\mathbf{p} - \mathbf{p}_0\|} \\ &= \lim_{(u,v,w)\to(1,-1,0)} \frac{\mathbf{F}(2u + v, v + w, u + w) - \mathbf{b} - A\mathbf{p}}{\sqrt{(u-1)^2 + (v+1)^2 + w^2}} = \mathbf{0}, \end{aligned}$$

the first order approximation theorem says that

$$\mathbf{b} + A\mathbf{p} = \mathbf{H}(\mathbf{p}_0) + \mathbf{DH}(\mathbf{p}_0)\,(\mathbf{p} - \mathbf{p}_0)\,.$$

Therefore,

$$A = \mathbf{DH}(\mathbf{p}_0) \quad \text{and} \quad \mathbf{b} = \mathbf{H}(\mathbf{p}_0) - A\mathbf{p}_0.$$

Notice that $\mathbf{G}(\mathbf{p}_0) = \mathbf{G}(1,-1,0) = (1,-1,1)$,

$$\mathbf{H}(\mathbf{p}_0) = \mathbf{H}(1,-1,0) = \mathbf{F}(1,-1,1) = (14,-1),$$

$$\mathbf{DG}(u,v,w) = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}, \qquad \mathbf{DF}(x,y,z) = \begin{bmatrix} 2x & 8y & 18z \\ yz & xz & xy \end{bmatrix}.$$

By chain rule,

$$\begin{aligned} A &= \mathbf{DF}(1,-1,1)\mathbf{DG}(1,-1,0) \\ &= \begin{bmatrix} 2 & -8 & 18 \\ -1 & 1 & -1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 22 & -6 & 10 \\ -3 & 0 & 0 \end{bmatrix}. \end{aligned}$$

It follows that

$$\mathbf{b} = \begin{bmatrix} 14 \\ -1 \end{bmatrix} - \begin{bmatrix} 22 & -6 & 10 \\ -3 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} = \begin{bmatrix} -14 \\ 2 \end{bmatrix}.$$

Example 4.23

Let α be a positive number, and let $f : \mathbb{R}^n \to \mathbb{R}$ be the function defined as

$$f(\mathbf{x}) = \|\mathbf{x}\|^\alpha.$$

Find the values of α so that f is differentiable.

Solution

Let $g : \mathbb{R}^n \to \mathbb{R}$ be the function

$$g(\mathbf{x}) = \|\mathbf{x}\|^2 = x_1^2 + x_2^2 + \cdots + x_n^2.$$

Then $g(\mathbb{R}^n) = [0,\infty)$, and $g(\mathbf{x}) = 0$ if and only if $\mathbf{x} = \mathbf{0}$.

Since g is a polynomial, it is infinitely differentiable. Let $h : [0, \infty) \to \mathbb{R}$ be the function $h(u) = u^{\alpha/2}$. Then h is differentiable on $(0, \infty)$. Since $f(\mathbf{x}) = (h \circ g)(\mathbf{x})$, chain rule implies that for all $\mathbf{x}_0 \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, f is differentiable at $\mathbf{x}_0$.

Now consider the point $\mathbf{x} = \mathbf{0}$. Notice that for $1 \leq i \leq n$, $f_{x_i}(\mathbf{0})$ exists provided that the limit

$$\lim_{h \to 0} \frac{f(h\mathbf{e}_i) - f(\mathbf{0})}{h} = \lim_{h \to 0} \frac{|h|^\alpha}{h}$$

exists. This is the case if $\alpha > 1$. Therefore, f is not differentiable at $\mathbf{x} = \mathbf{0}$ if $\alpha \leq 1$. If $\alpha > 1$, we find that $f_{x_i}(\mathbf{0}) = 0$ for all $1 \leq i \leq n$. Hence, $\nabla f(\mathbf{0}) = \mathbf{0}$. Since

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{f(\mathbf{h}) - f(\mathbf{0}) - \langle \nabla f(\mathbf{0}), \mathbf{h} \rangle}{\|\mathbf{h}\|} = \lim_{\mathbf{h} \to \mathbf{0}} \|\mathbf{h}\|^{\alpha - 1} = 0,$$

we conclude that when $\alpha > 1$, f is differentiable at $\mathbf{x} = \mathbf{0}$.

Therefore, f is differentiable if and only if $\alpha > 1$.

Example 4.24

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a twice continuously differentiable function, and let $g : \mathbb{R}^2 \to \mathbb{R}$ be the function defined as

$$g(r, \theta) = f(r \cos\theta, r \sin\theta).$$

Show that

$$\frac{\partial^2 g}{\partial r^2} + \frac{1}{r}\frac{\partial g}{\partial r} + \frac{1}{r^2}\frac{\partial^2 g}{\partial \theta^2} = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}.$$

Solution

Let $\mathbf{H} : \mathbb{R}^2 \to \mathbb{R}^2$ be the mapping defined by

$$\mathbf{H}(r, \theta) = (r \cos\theta, r \sin\theta).$$

Then $\mathbf{H}$ is infinitely differentiable, and $g = f \circ \mathbf{H}$. Let $x = H_1(r,\theta) = r\cos\theta$ and $y = H_2(r,\theta) = r\sin\theta$. By chain rule,

$$\begin{aligned}\frac{\partial g}{\partial r} &= \frac{\partial f}{\partial x}\frac{\partial x}{\partial r} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial r} = \cos\theta\frac{\partial f}{\partial x} + \sin\theta\frac{\partial f}{\partial y},\\ \frac{\partial g}{\partial \theta} &= \frac{\partial f}{\partial x}\frac{\partial x}{\partial \theta} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial \theta} = -r\sin\theta\frac{\partial f}{\partial x} + r\cos\theta\frac{\partial f}{\partial y}.\end{aligned}$$

Using product rule and chain rule, we then have

$$\frac{\partial^2 g}{\partial r^2} = \cos\theta\left(\frac{\partial^2 f}{\partial x^2}\frac{\partial x}{\partial r} + \frac{\partial^2 f}{\partial y\partial x}\frac{\partial y}{\partial r}\right) + \sin\theta\left(\frac{\partial^2 f}{\partial x\partial y}\frac{\partial x}{\partial r} + \frac{\partial^2 f}{\partial y^2}\frac{\partial y}{\partial r}\right).$$

Since f has continuous second order partial derivatives, $f_{xy} = f_{yx}$. Therefore,

$$\frac{\partial^2 g}{\partial r^2} = \cos^2\theta\frac{\partial^2 f}{\partial x^2} + 2\sin\theta\cos\theta\frac{\partial^2 f}{\partial x\partial y} + \sin^2\theta\frac{\partial^2 f}{\partial y^2}.$$

Similarly, we have

$$\begin{aligned}\frac{\partial^2 g}{\partial \theta^2} &= -r\sin\theta\left(\frac{\partial^2 f}{\partial x^2}\frac{\partial x}{\partial \theta} + \frac{\partial^2 f}{\partial y\partial x}\frac{\partial y}{\partial \theta}\right) + r\cos\theta\left(\frac{\partial^2 f}{\partial x\partial y}\frac{\partial x}{\partial \theta} + \frac{\partial^2 f}{\partial y^2}\frac{\partial y}{\partial \theta}\right)\\ &\quad - r\cos\theta\frac{\partial f}{\partial x} - r\sin\theta\frac{\partial f}{\partial y}\\ &= r^2\sin^2\theta\frac{\partial^2 f}{\partial x^2} - 2r^2\sin\theta\cos\theta\frac{\partial^2 f}{\partial x\partial y} + r^2\cos^2\theta\frac{\partial^2 f}{\partial y^2} - r\frac{\partial g}{\partial r}.\end{aligned}$$

From these, we obtain

$$\frac{\partial^2 g}{\partial r^2} + \frac{1}{r}\frac{\partial g}{\partial r} + \frac{1}{r^2}\frac{\partial^2 g}{\partial \theta^2} = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}.$$

Example 4.24 gives the *Laplacian*

$$\Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}$$

of f in polar coordinates. It is customary that one would abuse notation and write $g = f$, so that the formula takes the form

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = \frac{\partial^2 f}{\partial r^2} + \frac{1}{r}\frac{\partial f}{\partial r} + \frac{1}{r^2}\frac{\partial^2 f}{\partial \theta^2}.$$

Remark 4.13

We can use the chain rule to prove Theorem 4.13. Given that $\mathcal{O}$ is an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is a function that is differentiable at $\mathbf{x}_0$, we want to show that $\mathbf{D}_\mathbf{v}\mathbf{F}(\mathbf{x}_0)$ exists for any nonzero vector $\mathbf{v}$, and

$$\mathbf{D}_\mathbf{v}\mathbf{F}(\mathbf{x}_0) = \mathbf{DF}(\mathbf{x}_0)\mathbf{v}.$$

Since $\mathcal{O}$ is an open set that contains the point $\mathbf{x}_0$, there is an $r > 0$ such that $B(\mathbf{x}_0, r) \subset \mathcal{O}$. By definition,

$$\mathbf{D}_\mathbf{v}\mathbf{F}(\mathbf{x}_0) = \lim_{h\to 0} \frac{\mathbf{F}(\mathbf{x}_0 + h\mathbf{v}) - \mathbf{F}(\mathbf{x}_0)}{h} = \mathbf{g}'(0),$$

where $\mathbf{g} : (-r, r) \to \mathbb{R}^m$ is the function $\mathbf{g}(h) = \mathbf{F}(\mathbf{x}_0 + h\mathbf{v})$. Let $\boldsymbol{\gamma} : (-r, r) \to \mathbb{R}^n$ be the function defined as $\boldsymbol{\gamma}(h) = \mathbf{x}_0 + h\mathbf{v}$. Then $\boldsymbol{\gamma}$ is a differentiable function with $\boldsymbol{\gamma}'(h) = \mathbf{v}$. Since $\mathbf{g} = \mathbf{F} \circ \boldsymbol{\gamma}$, and $\boldsymbol{\gamma}(0) = \mathbf{x}_0$, the chain rule implies that $\mathbf{g}$ is differentiable at $h = 0$ and

$$\mathbf{g}'(0) = \mathbf{DF}(\mathbf{x}_0)\boldsymbol{\gamma}'(0) = \mathbf{DF}(\mathbf{x}_0)\mathbf{v}.$$

This completes the proof.

Definition 4.16 Tangent Line to a Curve

A curve in $\mathbb{R}^n$ is a continuous function $\boldsymbol{\gamma} : [a, b] \to \mathbb{R}^n$. Let c_0 be a point in (a, b). If the curve $\boldsymbol{\gamma}$ is differentiable at c_0, the tangent vector to the curve $\boldsymbol{\gamma}$ at the point $\boldsymbol{\gamma}(c_0)$ is the vector $\boldsymbol{\gamma}'(c_0)$ in $\mathbb{R}^n$, while the tangent line to the curve $\boldsymbol{\gamma}$ at the point $\boldsymbol{\gamma}(c_0)$ is the line in $\mathbb{R}^n$ given by $\mathbf{x} : \mathbb{R} \to \mathbb{R}^n$,

$$\mathbf{x}(t) = \boldsymbol{\gamma}(c_0) + t\boldsymbol{\gamma}'(c_0).$$

Remark 4.14 Tangent Lines and Tangent Planes

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $f : \mathcal{O} \to \mathbb{R}$ be a function that is differentiable at $\mathbf{x}_0$. We have seen that the tangent plane to the graph of f at the point $(\mathbf{x}_0, f(\mathbf{x}_0))$ has equation

$$x_{n+1} = f(\mathbf{x}_0) + \langle \nabla f(\mathbf{x}_0), \mathbf{x} - \mathbf{x}_0 \rangle.$$

Now assume that $r > 0$ and $\boldsymbol{\gamma} : (-r, r) \to \mathbb{R}^{n+1}$ is a differentiable curve in $\mathbb{R}^{n+1}$ that lies on the graph of f, and $\gamma(0) = (\mathbf{x}_0, f(\mathbf{x}_0))$. For all $t \in (-r, r)$,

$$\gamma_{n+1}(t) = f(\gamma_1(t), \ldots, \gamma_n(t)).$$

By chain rule, we find that

$$\gamma'_{n+1}(0) = \langle \nabla f(\mathbf{x}_0), \mathbf{v} \rangle, \qquad \text{where } \mathbf{v} = (\gamma'_1(0), \ldots, \gamma'_n(0)).$$

The vector $\mathbf{w} = (\mathbf{v}, \gamma'_{n+1}(0))$ is the tangent vector to the curve $\boldsymbol{\gamma}$ at the point $(\mathbf{x}_0, f(\mathbf{x}_0))$. The equation of the tangent line is

$$(x_1(t), \ldots, x_n(t), x_{n+1}(t)) = (\mathbf{x}_0, f(\mathbf{x}_0)) + t(\gamma'_1(0), \ldots, \gamma'_n(0), \gamma'_{n+1}(0)).$$

Thus, we find that

$$(x_1(t), \ldots, x_n(t)) = \mathbf{x}(t) = \mathbf{x}_0 + t\mathbf{v},$$

and

$$x_{n+1}(t) = f(\mathbf{x}_0) + t\gamma'_{n+1}(0).$$

These imply that

$$\begin{aligned} x_{n+1}(t) &= f(\mathbf{x}_0) + t\langle \nabla f(\mathbf{x}_0), \mathbf{v} \rangle \\ &= f(\mathbf{x}_0) + \langle \nabla f(\mathbf{x}_0), \mathbf{x}(t) - \mathbf{x}_0 \rangle. \end{aligned}$$

Thus, the tangent line to the curve $\boldsymbol{\gamma}$ lies in the tangent plane.
In fact, the tangent plane to the graph of a function f at a point can be characterized as the unique plane that contains all the tangent lines to the differentiable curves that lie on the graph and passing through that point.

Now we turn to the mean value theorem. For a single variable function, the mean value theorem says that given that $f : I \to \mathbb{R}$ is a differentiable function defined on the open interval I, if x_0 and $x_0 + h$ are two points in I, there exists

$c \in (0,1)$ such that

$$f(x_0 + h) - f(x_0) = hf'(x_0 + ch).$$

Notice that the point $x_0 + ch$ is a point strictly in between x_0 and $x_0 + h$. To generalize this theorem to multivariable functions, one natural question to ask is the following. If $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is a differentiable function defined on the open subset $\mathcal{O}$ of $\mathbb{R}^n$, $\mathbf{x}_0$ and $\mathbf{x}_0 + \mathbf{h}$ are points in $\mathcal{O}$ such that the line segment between them lies entirely in $\mathcal{O}$, does there exist a constant $c \in (0,1)$ such that

$$\mathbf{F}(\mathbf{x}_0 + \mathbf{h}) - \mathbf{F}(\mathbf{x}_0) = \mathbf{DF}(\mathbf{x}_0 + c\mathbf{h})\mathbf{h}?$$

When $m \geq 2$, the answer is no in general. Let us look at the following example.

Example 4.25

Consider the function $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$ defined as

$$\mathbf{F}(x, y) = (x^2 y, xy).$$

Show that there does not exist a contant $c \in (0,1)$ such that

$$\mathbf{F}(\mathbf{x}_0 + \mathbf{h}) - \mathbf{F}(\mathbf{x}_0) = \mathbf{DF}(\mathbf{x}_0 + c\mathbf{h})\mathbf{h},$$

when $\mathbf{x}_0 = (0,0)$ and $\mathbf{h} = (1,1)$.

Solution

Notice that

$$\mathbf{DF}(x, y) = \begin{bmatrix} 2xy & x^2 \\ y & x \end{bmatrix}.$$

When $\mathbf{x}_0 = (0,0)$ and $\mathbf{h} = (1,1)$, $\mathbf{x}_0 + c\mathbf{h} = (c, c)$. If there exists a constant $c \in (0,1)$ such that

$$\mathbf{F}(\mathbf{x}_0 + \mathbf{h}) - \mathbf{F}(\mathbf{x}_0) = \mathbf{DF}(\mathbf{x}_0 + c\mathbf{h})\mathbf{h},$$

then

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2c^2 & c^2 \\ c & c \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

This gives

$$3c^2 = 1 \quad \text{and} \quad 2c = 1.$$

But $2c = 1$ gives $c = 1/2$. When $c = 1/2$, $3c^2 = 3/4 \neq 1$. Hence, no such c can exist.

However, when $m = 1$, we indeed have a mean value theorem.

Theorem 4.16 The Mean Value Theorem

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $\mathbf{x}_0$ and $\mathbf{x}_0 + \mathbf{h}$ be two points in $\mathcal{O}$ such that the line segment between them lies entirely in $\mathcal{O}$. If $f : \mathcal{O} \to \mathbb{R}$ is a differentiable function, there exist a constant $c \in (0, 1)$ such that

$$f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) = \langle \nabla f(\mathbf{x}_0 + c\mathbf{h}), \mathbf{h} \rangle = \sum_{i=1}^{n} h_i \frac{\partial f}{\partial x_i}(\mathbf{x}_0 + c\mathbf{h}).$$

Proof

Define the function $\gamma : [0, 1] \to \mathbb{R}^n$ by $\gamma(t) = \mathbf{x}_0 + t\mathbf{h}$. Then γ is a differentiable function with $\gamma'(t) = \mathbf{h}$. Let $g = (f \circ \gamma) : [0, 1] \to \mathbb{R}$. Then

$$g(t) = (f \circ \gamma)(t) = f(\mathbf{x}_0 + t\mathbf{h}).$$

Since f and γ are differentiable, the chain rule implies that g is also differentiable and

$$g'(t) = \langle \nabla f(\mathbf{x}_0 + t\mathbf{h}), \gamma'(t) \rangle = \langle \nabla f(\mathbf{x}_0 + t\mathbf{h}), \mathbf{h} \rangle.$$

By mean value theorem for single variable functions, we find that there exists $c \in (0, 1)$ such that

$$g(1) - g(0) = g'(c).$$

In other words, there exists $c \in (0, 1)$ such that

$$f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) = \langle \nabla f(\mathbf{x}_0 + c\mathbf{h}), \mathbf{h} \rangle.$$

This completes the proof.

As in the single variable case, the mean value theorem has the following application.

Corollary 4.17

Let $\mathcal{O}$ be an open connected subset of $\mathbb{R}^n$, and let $f : \mathcal{O} \to \mathbb{R}$ be a function defined on $\mathcal{O}$. If f is differentiable and $\nabla f(\mathbf{x}) = \mathbf{0}$ for all $\mathbf{x} \in \mathcal{O}$, then f is a constant function.

Proof

If $\mathbf{u}$ and $\mathbf{v}$ are two points in $\mathcal{O}$ such that the line segment between them lies entirely in $\mathcal{O}$, then the mean value theorem implies that $f(\mathbf{u}) = f(\mathbf{v})$.
Since $\mathcal{O}$ is an open connected subset of $\mathbb{R}^n$, Theorem 3.16 says that any two points $\mathbf{u}$ and $\mathbf{v}$ in $\mathcal{O}$ can be joined by a polygonal path in $\mathcal{O}$. In other words, there are points $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_k$ in $\mathcal{O}$ such that $\mathbf{x}_0 = \mathbf{u}$, $\mathbf{x}_k = \mathbf{v}$, and for $1 \leq i \leq k$, the line segment between $\mathbf{x}_{i-1}$ and $\mathbf{x}_i$ lies entirely in $\mathcal{O}$. Therefore,

$$f(\mathbf{x}_{i-1}) = f(\mathbf{x}_i) \qquad \text{for all } 1 \leq i \leq k.$$

This proves that $f(\mathbf{u}) = f(\mathbf{v})$. Hence, f is a constant function.

Exercises 4.3

Question 1

Let $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^3$ be the function defined as

$$\mathbf{F}(x, y) = (x^2 + y^2, xy, x + y).$$

Find a vector $\mathbf{b}$ in $\mathbb{R}^3$ and a 3×2 matrix A such that

$$\lim_{(u,v)\to(1,-1)} \frac{\mathbf{F}(5u + 3v, u - 2v) - \mathbf{b} - A\mathbf{w}}{\sqrt{(u-1)^2 + (v+1)^2}} = \mathbf{0}, \quad \text{where } \mathbf{w} = \begin{bmatrix} u \\ v \end{bmatrix}.$$

Question 2

Let $\phi : \mathbb{R} \to \mathbb{R}$ and $\psi : \mathbb{R} \to \mathbb{R}$ be functions that have continuous second order derivatives, and let c be a constant. Define the function $f : \mathbb{R}^2 \to \mathbb{R}$ by

$$f(t, x) = \phi(x + ct) + \psi(x - ct).$$

Show that

$$\frac{\partial^2 f}{\partial t^2} - c^2 \frac{\partial^2 f}{\partial x^2} = 0.$$

Question 3

Let α be a constant, and let $f : \mathbb{R}^n \setminus \{\mathbf{0}\} \to \mathbb{R}$ be the function defined by

$$f(\mathbf{x}) = \|\mathbf{x}\|^\alpha.$$

Find the value(s) of α such that

$$\Delta f(\mathbf{x}) = \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2}(\mathbf{x}) = \frac{\partial^2 f}{\partial x_1^2}(\mathbf{x}) + \frac{\partial^2 f}{\partial x_2^2}(\mathbf{x}) + \cdots + \frac{\partial^2 f}{\partial x_n^2}(\mathbf{x}) = 0.$$

Question 4

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a function such that $f(0,0) = 2$ and

$$\frac{\partial f}{\partial x}(x,y) = 11 \quad \text{and} \quad \frac{\partial f}{\partial y} = -7 \qquad \text{for all } (x,y) \in \mathbb{R}^2.$$

Show that

$$f(x,y) = 2 + 11x - 7y \qquad \text{for all } (x,y) \in \mathbb{R}^2.$$

Question 5

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^2$, and let $u : \mathcal{O} \to \mathbb{R}$ and $v : \mathcal{O} \to \mathbb{R}$ be twice continuously differentiable functions. Define the function $\mathbf{F} : \mathcal{O} \to \mathbb{R}^2$ by

$$\mathbf{F}(x,y) = (u(x,y), v(x,y)).$$

Let $\mathcal{U}$ be an open subset of $\mathbb{R}^2$ that contains $\mathbf{F}(\mathcal{O})$, and let $f : \mathcal{U} \to \mathbb{R}$ be a twice continuously differentiable function. Define the function $g : \mathcal{O} \to \mathbb{R}$ by

$$g(x,y) = (f \circ \mathbf{F})(x,y) = f(u(x,y), v(x,y)).$$

Find g_{xx}, g_{xy} and g_{yy} in terms of the first and second order partial derivatives of u, v and f.

4.4 Second Order Approximations

In this section, we turn to consider second order approximations. We only consider a function $f : \mathcal{O} \to \mathbb{R}$ defined on an open subset $\mathcal{O}$ of $\mathbb{R}^n$ and whose codomain is $\mathbb{R}$. The function is said to be twice differentiable if it has first order partial derivatives, and each $f_{x_i} : \mathcal{O} \to \mathbb{R}$, $1 \leq i \leq n$, is a differentiable function. Notice that a twice differentiable function has continuous first order partial derivatives. Hence, it is differentiable. The differentiability of each f_{x_i}, $1 \leq i \leq n$ also implies that f has second order partial derivatives.

Lemma 4.18

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $f : \mathcal{O} \to \mathbb{R}$ be a twice differentiable function defined on $\mathcal{O}$. If $\mathbf{x}_0$ and $\mathbf{x}_0 + \mathbf{h}$ are two points in $\mathcal{O}$ such that the line segment between them lies entirely in $\mathcal{O}$, then there is a $c \in (0, 1)$ such that

$$\begin{aligned} f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) - \langle \nabla f(\mathbf{x}_0), \mathbf{h} \rangle &= \frac{1}{2} \mathbf{h}^T H_f(\mathbf{x}_0 + c\mathbf{h}) \mathbf{h} \\ &= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} h_i h_j \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0 + c\mathbf{h}). \end{aligned}$$

Proof

Given $\mathbf{x}_0 \in \mathcal{O}$, let r be a positive number such that $B(\mathbf{x}_0, r) \subset \mathcal{O}$. Define the function $g : (-r, r) \to \mathbb{R}$ by

$$g(t) = f(\mathbf{x}_0 + t\mathbf{h}).$$

Since $f : \mathcal{O} \to \mathbb{R}$ is differentiable, chain rule implies that $g : (-r, r) \to \mathbb{R}$ is differentiable and

$$g'(t) = \sum_{i=1}^{n} h_i \frac{\partial f}{\partial x_i}(\mathbf{x}_0 + t\mathbf{h}) = \langle \nabla f(\mathbf{x}_0 + t\mathbf{h}), \mathbf{h} \rangle.$$

Since each $f_{x_i} : \mathcal{O} \to \mathbb{R}$, $1 \leq i \leq n$ is differentiable, chain rule again implies that g' is differentiable and

$$g''(t) = \sum_{i=1}^{n}\sum_{j=1}^{n} h_i h_j \frac{\partial f}{\partial x_j \partial x_i}(\mathbf{x}_0 + t\mathbf{h}) = \mathbf{h}^T H_f(\mathbf{x}_0 + t\mathbf{h})\mathbf{h}.$$

By Lagrange's remainder theorem, there is a $c \in (0,1)$ such that

$$g(1) - g(0) - g'(0)(1-0) = \frac{g''(c)}{2}(1-0)^2.$$

This gives

$$f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) - \langle \nabla f(\mathbf{x}_0), \mathbf{h}\rangle = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} h_i h_j \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0 + c\mathbf{h}).$$

If a function has continuous second order partial derivatives, then it is twice differentiable, and Clairaut's theorem implies that its Hessian matrix is symmetric. For such a function, we can prove the second order approximation theorem.

Theorem 4.19 Second Order Approximation Theorem

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $f : \mathcal{O} \to \mathbb{R}$ be a twice continuously differentiable function defined on $\mathcal{O}$. We have the followings.

(a) $$\lim_{\mathbf{h}\to\mathbf{0}} \frac{f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) - \langle \nabla f(\mathbf{x}_0), \mathbf{h}\rangle - \frac{1}{2}\mathbf{h}^T H_f(\mathbf{x}_0)\mathbf{h}}{\|\mathbf{h}\|^2} = 0.$$

(b) If $Q(\mathbf{x})$ is a polynomial of degree at most two such that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{f(\mathbf{x}_0 + \mathbf{h}) - Q(\mathbf{x}_0 + \mathbf{h})}{\|\mathbf{h}\|^2} = 0,$$

then

$$Q(\mathbf{x}) = f(\mathbf{x}_0) + \langle \nabla f(\mathbf{x}_0), \mathbf{x} - \mathbf{x}_0\rangle + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T H_f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0). \quad (4.8)$$

Combining (a) and (b), the second order approximation theorem says that for a twice continuously differentiable function, there exists a unique polynomial of degree at most 2 which is a second order approximation of the function.

Proof

Let us prove part (a) first. Since $\mathcal{O}$ is open, there is an $r > 0$ such that $B(\mathbf{x}_0, r) \subset \mathcal{O}$. For each $\mathbf{h}$ in $\mathbb{R}^n$ with $\|\mathbf{h}\| < r$, Lemma 4.18 says that there is a $c_{\mathbf{h}} \in (0,1)$ such that

$$f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) - \langle \nabla f(\mathbf{x}_0), \mathbf{h} \rangle = \frac{1}{2} \mathbf{h}^T H_f(\mathbf{x}_0 + c\mathbf{h}) \mathbf{h}.$$

Therefore, if $0 < \|\mathbf{h}\| < r$,

$$\begin{aligned}
&\left| \frac{f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) - \langle \nabla f(\mathbf{x}_0), \mathbf{h} \rangle - \frac{1}{2} \mathbf{h}^T H_f(\mathbf{x}_0) \mathbf{h}}{\|\mathbf{h}\|^2} \right| \\
&= \frac{1}{2} \left| \sum_{i=1}^n \sum_{j=1}^n \frac{h_i h_j}{\|\mathbf{h}\|^2} \left(\frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0 + c_{\mathbf{h}} \mathbf{h}) - \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0) \right) \right| \\
&\leq \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \frac{|h_i||h_j|}{\|\mathbf{h}\|^2} \left| \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0 + c_{\mathbf{h}} \mathbf{h}) - \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0) \right| \\
&\leq \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \left| \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0 + c_{\mathbf{h}} \mathbf{h}) - \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0) \right|.
\end{aligned}$$

Since $c_{\mathbf{h}} \in (0,1)$, $\lim\limits_{\mathbf{h} \to \mathbf{0}} (\mathbf{x}_0 + c_{\mathbf{h}} \mathbf{h}) = \mathbf{x}_0$. For all $1 \leq i \leq n$, $1 \leq j \leq n$, $f_{x_j x_i}$ is continuous. Hence,

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0 + c_{\mathbf{h}} \mathbf{h}) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0).$$

This proves that

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) - \langle \nabla f(\mathbf{x}_0), \mathbf{h} \rangle - \frac{1}{2} \mathbf{h}^T H_f(\mathbf{x}_0) \mathbf{h}}{\|\mathbf{h}\|^2} = 0.$$

To prove part (b), let

$$P(\mathbf{x}) = f(\mathbf{x}_0) + \langle \nabla f(\mathbf{x}_0), \mathbf{x} - \mathbf{x}_0 \rangle + \frac{1}{2} (\mathbf{x} - \mathbf{x}_0)^T H_f(\mathbf{x}_0) (\mathbf{x} - \mathbf{x}_0).$$

Part (a) says that

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{f(\mathbf{x}_0 + \mathbf{h}) - P(\mathbf{x}_0 + \mathbf{h})}{\|\mathbf{h}\|^2} = 0. \tag{4.9}$$

Since $Q(\mathbf{x})$ is a polynomial of degree at most two in $\mathbf{x}$, $Q(\mathbf{x}_0 + \mathbf{h})$ is a polynomial of degree at most two in $\mathbf{h}$. Therefore, we can write $Q(\mathbf{x}_0 + \mathbf{h})$ as

$$Q(\mathbf{x}_0 + \mathbf{h}) = c + \sum_{i=1}^{n} b_i h_i + \frac{1}{2}\sum_{i=1}^{n} a_{ii} h_i^2 + \sum_{1 \le i < j \le n} a_{ij} h_i h_j.$$

Since

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{f(\mathbf{x}_0 + \mathbf{h}) - Q(\mathbf{x}_0 + \mathbf{h})}{\|\mathbf{h}\|^2} = 0,$$

subtracting (4.9) gives

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{P(\mathbf{x}_0 + \mathbf{h}) - Q(\mathbf{x}_0 + \mathbf{h})}{\|\mathbf{h}\|^2} = 0. \tag{4.10}$$

It follows that

$$\lim_{\mathbf{h}\to\mathbf{0}} \left(P(\mathbf{x}_0 + \mathbf{h}) - Q(\mathbf{x}_0 + \mathbf{h})\right) = 0, \tag{4.11}$$

and

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{P(\mathbf{x}_0 + \mathbf{h}) - Q(\mathbf{x}_0 + \mathbf{h})}{\|\mathbf{h}\|} = 0. \tag{4.12}$$

Since f has continuous second order partial derivatives, $f_{x_j x_i}(\mathbf{x}_0) = f_{x_i x_j}(\mathbf{x}_0)$. Thus,

$$\begin{aligned}
&P(\mathbf{x}_0 + \mathbf{h}) - Q(\mathbf{x}_0 + \mathbf{h}) \\
&= (f(\mathbf{x}_0) - c) + \sum_{i=1}^{n} h_i \left(\frac{\partial f}{\partial x_i}(\mathbf{x}_0) - b_i\right) \\
&\quad + \frac{1}{2}\sum_{i=1}^{n} h_i^2 \left(\frac{\partial^2 f}{\partial x_i^2}(\mathbf{x}_0) - a_{ii}\right) + \sum_{1 \le i < j \le n} h_i h_j \left(\frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0) - a_{ij}\right).
\end{aligned}$$

Eq. (4.11) implies that $c = f(\mathbf{x}_0)$. Then eq. (4.12) implies that

$$b_i = \frac{\partial f}{\partial x_i}(\mathbf{x}_0) \qquad \text{for all } 1 \le i \le n.$$

Finally, (4.10) implies that for any $1 \le i \le j \le n$,

$$a_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}_0).$$

This completes the proof that $Q(\mathbf{x}) = P(\mathbf{x})$.

Example 4.26

Find a polynomial $Q(x, y)$ of degree at most 2 such that

$$\lim_{(x,y)\to(1,2)} \frac{\sin(4x^2 - y^2) - Q(x, y)}{(x - 1)^2 + (y - 2)^2} = 0.$$

Solution

Since $g(x, y) = 4x^2 - y^2$ is a polynomial function, it is infinitely differentiable. Since the sine function is also infinitely differentiable, the function $f(x, y) = \sin(4x^2 - y^2)$ is infinitely differentiable.

$$\begin{aligned} f_x(x, y) &= 8x\cos(4x^2 - y^2), \qquad f_y(x, y) = -2y\cos(4x^2 - y^2), \\ f_{xx}(x, y) &= 8\cos(4x^2 - y^2) - 64x^2\sin(4x^2 - y^2), \\ f_{xy}(x, y) &= f_{yx}(x, y) = 16xy\sin(4x^2 - y^2), \\ f_{yy}(x, y) &= -2\cos(4x^2 - y^2) - 4y^2\sin(4x^2 - y^2). \end{aligned}$$

Hence,

$$f(1, 2) = 0, \quad f_x(1, 2) = 8, \quad f_y(1, 2) = -4,$$
$$f_{xx}(1, 2) = 8, \quad f_{xy}(1, 2) = 0, \quad f_{yy}(1, 2) = -2.$$

By the second order approximation theorem,

$$\begin{aligned} Q(x, y) &= f(1, 2) + f_x(1, 2)(x - 1) + f_y(1, 2)(y - 2) + \frac{1}{2}f_{xx}(1, 2)(x - 1)^2 \\ &\quad + f_{xy}(1, 2)(x - 1)(y - 2) + \frac{1}{2}f_{yy}(1, 2)(y - 2)^2 \\ &= 8(x - 1) - 4(y - 2) + 4(x - 1)^2 - (y - 2)^2 \\ &= 4x^2 - y^2. \end{aligned}$$

Example 4.27

Determine whether the limit $\lim\limits_{(x,y)\to(0,0)} \dfrac{e^{x+y} - 1 - x - y}{x^2 + y^2}$ exists. If yes, find the limit.

Solution

Since the exponential funtion and the function $g(x,y) = x+y$ are infinitely differentiable, the function $f(x,y) = e^{x+y}$ is infinitely differentiable. By the second order approximation theorem,

$$\lim_{(x,y)\to(0,0)} \frac{f(x,y) - Q(x,y)}{x^2+y^2} = 0,$$

where

$$\begin{aligned} Q(x,y) = f(0,0) &+ x\frac{\partial f}{\partial x}(0,0) + y\frac{\partial f}{\partial y}(0,0) \\ &+ \frac{1}{2}x^2\frac{\partial^2 f}{\partial x^2}(0,0) + xy\frac{\partial^2 f}{\partial x \partial y}(0,0) + \frac{1}{2}y^2\frac{\partial^2 f}{\partial y^2}(0,0). \end{aligned}$$

Now

$$\frac{\partial f}{\partial x}(x,y) = \frac{\partial f}{\partial y}(x,y) = \frac{\partial^2 f}{\partial x^2}(x,y) = \frac{\partial^2 f}{\partial x \partial y}(x,y) = \frac{\partial^2 f}{\partial y^2}(x,y) = e^{x+y}.$$

Thus,

$$f(0,0) = \frac{\partial f}{\partial x}(0,0) = \frac{\partial f}{\partial y}(0,0) = \frac{\partial^2 f}{\partial x^2}(0,0) = \frac{\partial^2 f}{\partial x \partial y}(0,0) = \frac{\partial^2 f}{\partial y^2}(0,0) = 1.$$

It follows that

$$Q(x,y) = 1 + x + y + \frac{1}{2}x^2 + xy + \frac{1}{2}y^2.$$

Hence,

$$\lim_{(x,y)\to(0,0)} \frac{e^{x+y} - 1 - x - y - \frac{1}{2}x^2 - xy - \frac{1}{2}y^2}{x^2+y^2} = 0. \tag{4.13}$$

If

$$\lim_{(x,y)\to(0,0)} \frac{e^{x+y} - 1 - x - y}{x^2+y^2} = a$$

exists, subtracting (4.13) shows that

$$a = \lim_{(x,y)\to(0,0)} h(x,y), \qquad \text{where } h(x,y) = \frac{\frac{1}{2}x^2 + xy + \frac{1}{2}y^2}{x^2+y^2}.$$

This implies that if $\{\mathbf{w}_k\}$ is a sequence in $\mathbb{R}^2 \setminus \{\mathbf{0}\}$ that converges to $(0,0)$, then the sequence $\{h(\mathbf{w}_k)\}$ converges to a. For $k \in \mathbb{Z}^+$, let

$$\mathbf{u}_k = \left(\frac{1}{k}, 0\right), \quad \mathbf{v}_k = \left(\frac{1}{k}, \frac{1}{k}\right).$$

Then $\{\mathbf{u}_k\}$ and $\{\mathbf{v}_k\}$ are sequences in $\mathbb{R}^2 \setminus \{\mathbf{0}\}$ that converge to $(0,0)$. Hence, the sequences $\{h(\mathbf{u}_k)\}$ and $\{h(\mathbf{v}_k)\}$ both converge to a. Since

$$h(\mathbf{u}_k) = \frac{1}{2}, \quad h(\mathbf{v}_k) = 1 \qquad \text{for all } k \in \mathbb{Z}^+,$$

the sequence $\{h(\mathbf{u}_k)\}$ converges to $\frac{1}{2}$, while the sequence $\{h(\mathbf{v}_k)\}$ converges to 1. This gives a contradiction. Hence, the limit

$$\lim_{(x,y)\to(0,0)} \frac{e^{x+y} - 1 - x - y}{x^2 + y^2}$$

does not exist.

Exercises 4.4

Question 1

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function

$$f(x, y) = x^2y + 4xy^2.$$

Find a polynomial $Q(x, y)$ of degree at most 2 such that

$$\lim_{(x,y)\to(1,-1)} \frac{f(x, y) - Q(x, y)}{(x - 1)^2 + (y + 1)^2} = 0.$$

Question 2

Determine whether the limit $\lim\limits_{(x,y)\to(0,0)} \dfrac{\sin(x + y) - x - y}{x^2 + y^2}$ exists. If yes, find the limit.

Question 3

Determine whether the limit $\lim\limits_{(x,y)\to(0,0)} \dfrac{\cos(x + y) - 1}{x^2 + y^2}$ exists. If yes, find the limit.

4.5 Local Extrema

In this section, we use differential calculus to study local extrema of a function $f : \mathcal{O} \to \mathbb{R}$ that is defined on an open subset $\mathcal{O}$ of $\mathbb{R}^n$. The definition of local extrema that we give here is only restricted to such functions.

Definition 4.17 Local Maximum and Local Minimum

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $f : \mathcal{O} \to \mathbb{R}$ be a function defined on $\mathcal{O}$.

1. The point $\mathbf{x}_0$ is called a *local maximizer* of f provided that there is a $\delta > 0$ such that $B(\mathbf{x}_0, \delta) \subset \mathcal{O}$ and for all $\mathbf{x} \in B(\mathbf{x}_0, \delta)$,

$$f(\mathbf{x}) \leq f(\mathbf{x}_0).$$

   The value $f(\mathbf{x}_0)$ is called a local maximum value of f.

2. The point $\mathbf{x}_0$ is called a *local minimizer* of f provided that there is a $\delta > 0$ such that $B(\mathbf{x}_0, \delta) \subset \mathcal{O}$ and for all $\mathbf{x} \in B(\mathbf{x}_0, \delta)$,

$$f(\mathbf{x}) \geq f(\mathbf{x}_0).$$

   The value $f(\mathbf{x}_0)$ is called a local minimum value of f.

3. The point $\mathbf{x}_0$ is called a local extremizer if it is either a local maximizer or a local minimizer. The value $f(\mathbf{x}_0)$ is called a local extreme value if it is either a local maximum value or a local minimum value.

From the definition, it is obvious that $\mathbf{x}_0$ is a local minimizer of the function $f : \mathcal{O} \to \mathbb{R}$ if and only if it is a local maximizer of the function $-f : \mathcal{O} \to \mathbb{R}$.

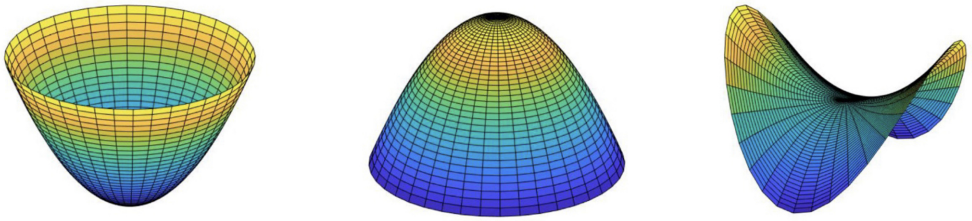
Example 4.28

(a) For the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = x^2 + y^2$, $(0, 0)$ is a local minimizer.

(b) For the function $g : \mathbb{R}^2 \to \mathbb{R}$, $g(x, y) = -x^2 - y^2$, $(0, 0)$ is a local maximizer.

(c) For the function $h : \mathbb{R}^2 \to \mathbb{R}$, $h(x, y) = x^2 - y^2$, $\mathbf{0} = (0, 0)$ is neither a local maximizer nor a local minimizer. For any $\delta > 0$, let $r = \delta/2$. The points $\mathbf{u} = (r, 0)$ and $\mathbf{v} = (0, r)$ are in $B(\mathbf{0}, \delta)$, but

$$h(\mathbf{u}) = r^2 > 0 = h(\mathbf{0}), \qquad h(\mathbf{v}) = -r^2 < 0 = h(\mathbf{0}).$$



$z = f(x, y)$ $z = g(x, y)$ $z = h(x, y)$

Figure 4.10: The functions $f(x, y)$, $g(x, y)$ and $h(x, y)$ defined in Example 4.28.

The following theorem gives a necessary condition for a point to be a local extremum if the function has partial derivatives at that point.

Theorem 4.20

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $f : \mathcal{O} \to \mathbb{R}$ be a function defined on $\mathcal{O}$. If $\mathbf{x}_0$ is a local extremizer and f has partial derivatives at $\mathbf{x}_0$, then the gradient of f at $\mathbf{x}_0$ is the zero vector, namely, $\nabla f(\mathbf{x}_0) = \mathbf{0}$.

Proof

Without loss of generality, assume that $\mathbf{x}_0$ is a local minimizer. Then there is a $\delta > 0$ such that $B(\mathbf{x}_0, \delta) \subset \mathcal{O}$ and

$$f(\mathbf{x}) \geq f(\mathbf{x}_0) \qquad \text{for all } \mathbf{x} \in B(\mathbf{x}_0, \delta). \tag{4.14}$$

For $1 \leq i \leq n$, consider the function $g_i : (-\delta, \delta) \to \mathbb{R}$ defined by $g_i(t) = f(\mathbf{x}_0 + t\mathbf{e}_i)$. By the definition of partial derivatives, g_i is differentiable at $t = 0$ and

$$g_i'(0) = \frac{\partial f}{\partial x_i}(\mathbf{x}_0).$$

Eq. (4.14) implies that

$$g_i(t) \geq g_i(0) \qquad \text{for all } t \in (-\delta, \delta).$$

In other words, $t = 0$ is a local minimizer of the function $g_i : (-\delta, \delta) \to \mathbb{R}$. From the theory of single variable analysis, we must have $g_i'(0) = 0$. Hence, $f_{x_i}(\mathbf{x}_0) = 0$ for all $1 \leq i \leq n$. This proves that $\nabla f(\mathbf{x}_0) = \mathbf{0}$.

Theorem 4.20 prompts us to make the following definition.

Definition 4.18 Stationary Points

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $f : \mathcal{O} \to \mathbb{R}$ be a function defined on $\mathcal{O}$. If f has partial derivatives at $\mathbf{x}_0$ and $\nabla f(\mathbf{x}_0) = \mathbf{0}$, we call $\mathbf{x}_0$ a stationary point of f.

Theorem 4.20 says that if $f : \mathcal{O} \to \mathbb{R}$ has partial derivatives at $\mathbf{x}_0$, a necessary condition for $\mathbf{x}_0$ to be a local extremizer is that it is a stationary point.

Example 4.29

For all the three functions f, g and h defined in Example 4.28, the point $\mathbf{0} = (0, 0)$ is a stationary point. However, $\mathbf{0}$ is local minimizer of f, a local maximizer of g, but neither a local maximizer nor a local minimizer of h.

The behavior of the function $h(x, y) = x^2 - y^2$ in Example 4.28 prompts us to make the following definition.

Definition 4.19 Saddle Points

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $f : \mathcal{O} \to \mathbb{R}$ be a function defined on $\mathcal{O}$. The point $\mathbf{x}_0$ is a saddle point of the function f if it is a stationary point of f, but it is not a local extremizer. In other words, $\nabla f(\mathbf{x}_0) = \mathbf{0}$, but for any $\delta > 0$, there exist $\mathbf{x}_1$ and $\mathbf{x}_2$ in $B(\mathbf{x}_0, \delta) \cap \mathcal{O}$ such that

$$f(\mathbf{x}_1) > f(\mathbf{x}_0) \quad \text{and} \quad f(\mathbf{x}_2) < f(\mathbf{x}_0).$$

Example 4.30

$(0, 0)$ is a saddle point of the function $h : \mathbb{R}^2 \to \mathbb{R}$, $h(x, y) = x^2 - y^2$.

By definition, if $\mathbf{x}_0$ is a stationary point of the function $f : \mathcal{O} \to \mathbb{R}$, then it is either a local maximizer, a local minimizer, or a saddle point. If $f : \mathcal{O} \to \mathbb{R}$ has continuous second order partial derivatives at $\mathbf{x}_0$, we can use the second derivative test to partially determine whether $\mathbf{x}_0$ is a local maximizer, a local minimizer, or a saddle point. When $n = 1$, we have seen that a stationary point x_0 of a function f is a local minimum if $f''(x_0) > 0$. It is a local maximum if $f''(x_0) < 0$. For multivariable functions, it is natural to expect that whether $\mathbf{x}_0$ is a local extremizer depends on the definiteness of the Hessian matrix $H_f(\mathbf{x}_0)$.

In Section 2.1, we have discussed the classification of a symmetric matrix. It is either positive semi-definite, negative semi-definite or indefinite. Among the positive semi-definite ones, there are those that are positive definite. Among the negative semi-definite matrices, there are those which are negative definite.

Theorem 4.21 Second Derivative Test

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $f : \mathcal{O} \to \mathbb{R}$ be a twice continuously differentiable function defined on $\mathcal{O}$. Assume that $\mathbf{x}_0$ is a stationary point of $f : \mathcal{O} \to \mathbb{R}$.

(i) If $H_f(\mathbf{x}_0)$ is positive definite, then $\mathbf{x}_0$ is a local minimizer of f.

(ii) If $H_f(\mathbf{x}_0)$ is negative definite, then $\mathbf{x}_0$ is a local maximizer of f.

(iii) If $H_f(\mathbf{x}_0)$ is indefinite, then $\mathbf{x}_0$ is a saddle point.

The cases that are not covered in the second derivative test are the cases where $H_f(\mathbf{x}_0)$ is positive semi-definite but not positive definite, or $H_f(\mathbf{x}_0)$ is negative semi-definite but not negative definite. These are the inconclusive cases.

Proof of the Second Derivative Test

Notice that (i) and (ii) are equivalent since $\mathbf{x}_0$ is a local minimizer of f if and only if it is a local maximizer of $-f$, and $H_{-f} = -H_f$. A symmetric matrix A is positive definite if and only if $-A$ is negative definite. Thus, we only need to prove (i) and (iii).

Since $\mathbf{x}_0$ is a stationary point, $\nabla f(\mathbf{x}_0) = \mathbf{0}$. It follows from the second order approximation theorem that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{f(\mathbf{x}_0+\mathbf{h}) - f(\mathbf{x}_0) - \frac{1}{2}\mathbf{h}^T H_f(\mathbf{x}_0)\mathbf{h}}{\|\mathbf{h}\|^2} = 0. \tag{4.15}$$

To prove (i), asume that $H_f(\mathbf{x}_0)$ is positive definite. By Theorem 2.9, there is a positive number c such that

$$\mathbf{h}^T H_f(\mathbf{x}_0)\mathbf{h} \geq c\|\mathbf{h}\|^2 \qquad \text{for all } \mathbf{h} \in \mathbb{R}^n.$$

Eq. 4.15 implies that there is a $\delta > 0$ such that $B(\mathbf{x}_0, \delta) \subset \mathcal{O}$ and for all $\mathbf{h}$ with $0 < \|\mathbf{h}\| < \delta$,

$$\left| \frac{f(\mathbf{x}_0+\mathbf{h}) - f(\mathbf{x}_0) - \frac{1}{2}\mathbf{h}^T H_f(\mathbf{x}_0)\mathbf{h}}{\|\mathbf{h}\|^2} \right| < \frac{c}{3}.$$

Therefore,

$$\left| f(\mathbf{x}_0+\mathbf{h}) - f(\mathbf{x}_0) - \frac{1}{2}\mathbf{h}^T H_f(\mathbf{x}_0)\mathbf{h} \right| \leq \frac{c}{3}\|\mathbf{h}\|^2 \qquad \text{for all } \|\mathbf{h}\| < \delta.$$

This implies that for all $\mathbf{h}$ with $\|\mathbf{h}\| < \delta$,

$$f(\mathbf{x}_0+\mathbf{h}) - f(\mathbf{x}_0) \geq \frac{1}{2}\mathbf{h}^T H_f(\mathbf{x}_0)\mathbf{h} - \frac{c}{3}\|\mathbf{h}\|^2 \geq \frac{c}{6}\|\mathbf{h}\|^2 \geq 0.$$

Thus, $f(\mathbf{x}) \geq f(\mathbf{x}_0)$ for all $\mathbf{x} \in B(\mathbf{x}_0, \delta)$. This shows that $\mathbf{x}_0$ is a local minimizer of f.

Now to prove (iii), assume that $H_f(\mathbf{x}_0)$ is indefinite. Then there exist unit vectors $\mathbf{u}_1$ and $\mathbf{u}_2$ so that

$$\varepsilon_1 = \mathbf{u}_1^T H_f(\mathbf{x}_0)\mathbf{u}_1 < 0, \qquad \varepsilon_2 = \mathbf{u}_2^T H_f(\mathbf{x}_0)\mathbf{u}_2 > 0.$$

Let $\varepsilon = \frac{1}{2}\min\{|\varepsilon_1|, \varepsilon_2\}$. Eq. (4.15) implies that there is a $\delta_0 > 0$ such that $B(\mathbf{x}_0, \delta_0) \subset \mathcal{O}$ and for all $\mathbf{h}$ with $0 < \|\mathbf{h}\| < \delta_0$,

$$\left| f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) - \frac{1}{2}\mathbf{h}^T H_f(\mathbf{x}_0)\mathbf{h} \right| < \varepsilon\|\mathbf{h}\|^2. \tag{4.16}$$

For any $\delta > 0$, let $r = \frac{1}{2}\min\{\delta, \delta_0\}$. Then the points $\mathbf{x}_1 = \mathbf{x}_0 + r\mathbf{u}_1$ and $\mathbf{x}_2 = \mathbf{x}_0 + r\mathbf{u}_2$ are in the ball $B(\mathbf{x}_0, \delta)$ and the ball $B(\mathbf{x}_0, \delta_0)$. Eq. (4.16) implies that for $i = 1, 2$,

$$-r^2\varepsilon \le f(\mathbf{x}_0 + r\mathbf{u}_i) - f(\mathbf{x}_0) - \frac{r^2}{2}\mathbf{u}_i^T H_f(\mathbf{x}_0)\mathbf{u}_i < r^2\varepsilon.$$

Therefore,

$$f(\mathbf{x}_0 + r\mathbf{u}_1) - f(\mathbf{x}_0) < r^2\left(\frac{1}{2}\mathbf{u}_1^T H_f(\mathbf{x}_0)\mathbf{u}_1 + \varepsilon\right) = r^2\left(\frac{1}{2}\varepsilon_1 + \varepsilon\right) \le 0$$

since $\varepsilon \le -\frac{1}{2}\varepsilon_1$; while

$$f(\mathbf{x}_0 + r\mathbf{u}_2) - f(\mathbf{x}_0) > r^2\left(\frac{1}{2}\mathbf{u}_2^T H_f(\mathbf{x}_0)\mathbf{u}_2 - \varepsilon\right) = r^2\left(\frac{1}{2}\varepsilon_2 - \varepsilon\right) \ge 0$$

since $\varepsilon \le \frac{1}{2}\varepsilon_2$. Thus, $\mathbf{x}_1$ and $\mathbf{x}_2$ are points in $B(\mathbf{x}_0, \delta)$, but $f(\mathbf{x}_1) < f(\mathbf{x}_0)$ while $f(\mathbf{x}_2) > f(\mathbf{x}_0)$. These show that $\mathbf{x}_0$ is a saddle point.

A symmetric matrix is positive definite if and only if all its eigenvalues are positive. It is negative definite if and only if all its eigenvalues are negative. It is indefinite if it has at least one positive eigenvalue, and at least one negative eigenvalue. For a diagonal matrix, its eigenvalues are the entries on the diagonal.

Let us revisit Example 4.28.

Example 4.31

For the functions considered in Example 4.28, we have seen that $(0,0)$ is a stationary point of each of them. Notice that $H_f(0,0) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$ is positive definite, $H_g(0,0) = \begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix}$ is negative definite, $H_h(0,0) = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$ is indefinite. Therefore, $(0,0)$ is a local minimizer of f, a local maximizer of g, and a saddle point of h.

Now let us look at an example which shows that when the Hessian matrix is positive semi-definite but not positive definite, we cannot make any conclusion about the nature of a stationary point.

Example 4.32

Consider the functions $f : \mathbb{R}^2 \to \mathbb{R}$ and $g : \mathbb{R}^2 \to \mathbb{R}$ given respectively by

$$f(x,y) = x^2 + y^4, \quad g(x,y) = x^2 - y^4.$$

These are infinitely differentiable functions. It is easy to check that $(0,0)$ is a stationary point of both of them. Now,

$$H_f(0,0) = H_g(0,0) = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$$

is a positive semi-definite matrix. However, $(0,0)$ is a local minimizer of f, but a saddle point of g.

To determine the definiteness of an $n \times n$ symmetric matrix by looking at the sign of its eigenvalues is ineffective when $n \geq 3$. There is an easier way to determine whether a symmetric matrix is positive definite. Let us first introduce the definition of principal submatrices.

Definition 4.20 Principal Submatrices

Let A be an $n \times n$ matrix. For $1 \leq k \leq n$, the k^{th}-principal submatrix M_k of A is the $k \times k$ matrix consists of the first k rows and first k columns of A.

Example 4.33

For the matrix $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$, the first, second and third principal submatrices are

$$M_1 = \begin{bmatrix} 1 \end{bmatrix}, \; M_2 = \begin{bmatrix} 1 & 2 \\ 4 & 5 \end{bmatrix}, \; M_3 = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

respectively.

Theorem 4.22 Sylvester's Criterion for Positive Definiteness

An $n \times n$ symmetric matrix A is positive definite if and only if $\det M_k > 0$ for all $1 \leq k \leq n$, where M_k is its k^{th} principal submatrix.

The proof of this theorem is given in Appendix A. Using the fact that a symmetric matrix A is negative definite if and only if $-A$ is positive definite, it is easy to obtain a criterion for a symmetric matrix to be negative definite in terms of the determinants of its principal submatrices.

Theorem 4.23 Sylvester's Criterion for Negative Definiteness

An $n \times n$ symmetric matrix A is negative definite if and only if $(-1)^k \det M_k > 0$ for all $1 \leq k \leq n$, where M_k is its k^{th} principal submatrix.

Example 4.34

Consider the matrix

$$A = \begin{bmatrix} 1 & 2 & -3 \\ -1 & 4 & 2 \\ -3 & 5 & 8 \end{bmatrix}.$$

Since

$$\det M_1 = 1,\ \det M_2 = 6,\ \det M_3 = \det A = 5$$

are all positive, A is positive definite.

For a function $f : \mathcal{O} \to \mathbb{R}$ defined on an open subset $\mathcal{O}$ of $\mathbb{R}^2$, we have the following.

Theorem 4.24

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^2$. Suppose that (x_0, y_0) is a stationary point of the twice continuously differentiable function $f : \mathcal{O} \to \mathbb{R}$. Let

$$D(x_0, y_0) = \frac{\partial^2 f}{\partial x^2}(x_0, y_0)\frac{\partial^2 f}{\partial y^2}(x_0, y_0) - \left[\frac{\partial^2 f}{\partial x \partial y}(x_0, y_0)\right]^2.$$

(i) If $\dfrac{\partial^2 f}{\partial x^2}(x_0, y_0) > 0$ and $D(x_0, y_0) > 0$, then the point (x_0, y_0) is a local minimizer of f.

(ii) If $\dfrac{\partial^2 f}{\partial x^2}(x_0, y_0) < 0$ and $D(x_0, y_0) > 0$, then the point (x_0, y_0) is a local maximizer of f.

(iii) If $D(x_0, y_0) < 0$, the point (x_0, y_0) is a saddle point of f.

Proof

We notice that

$$H_f(x_0, y_0) = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x^2}(x_0, y_0) & \dfrac{\partial^2 f}{\partial x \partial y}(x_0, y_0) \\ \dfrac{\partial^2 f}{\partial x \partial y}(x_0, y_0) & \dfrac{\partial^2 f}{\partial y^2}(x_0, y_0) \end{bmatrix}.$$

Hence, $\dfrac{\partial^2 f}{\partial x^2}(x_0, y_0)$ is the determinant of the first principal submatrix of $H_f(x_0, y_0)$, while $D(x_0, y_0)$ is the determinant of $H_f(x_0, y_0)$, the second principal submatrix of $H_f(x_0, y_0)$. Thus, (i) and (ii) follow from the Sylvester criteria as well as the second derivative test.
For (iii), we notice that the 2×2 matrix $H_f(x_0, y_0)$ is indefinite if and only if it has one positive eigenvalue and one negative eigenvalue, if and only if $D(x_0, y_0) = \det H_f(x_0, y_0) < 0$.

Now we look at some examples of the applications of the second derivative test.

Example 4.35

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function defined as

$$f(x, y) = x^4 + y^4 + 4xy.$$

Find the stationary points of f and classify them.

Solution

Since f is a polynomial function, it is infinitely differentiable.

$$\nabla f(x, y) = (4x^3 + 4y, 4y^3 + 4x).$$

To find the stationary points, we need to solve the system of equations

$$\begin{cases} x^3 + y = 0 \\ y^3 + x = 0 \end{cases}.$$

From the first equation, we have $y = -x^3$. Substitute into the second equation gives

$$-x^9 + x = 0,$$

or equivalently,

$$x(x^8 - 1) = 0.$$

Thus, $x = 0$ or $x = \pm 1$. When $x = 0$, $y = 0$. When $x = \pm 1$, $y = \mp 1$. Therefore, the stationary points of f are $\mathbf{u}_1 = (0, 0)$, $\mathbf{u}_2 = (1, -1)$ and $\mathbf{u}_3 = (-1, 1)$. Now,

$$H_f(x, y) = \begin{bmatrix} 12x^2 & 4 \\ 4 & 12y^2 \end{bmatrix}.$$

Therefore,

$$H_f(\mathbf{u}_1) = \begin{bmatrix} 0 & 4 \\ 4 & 0 \end{bmatrix}, \quad H_f(\mathbf{u}_2) = H_f(\mathbf{u}_3) = \begin{bmatrix} 12 & 4 \\ 4 & 12 \end{bmatrix}.$$

It follows that

$$D(\mathbf{u}_1) = -16 < 0, \quad D(\mathbf{u}_2) = D(\mathbf{u}_3) = 128 > 0.$$

Since $f_{xx}(\mathbf{u}_2) = f_{xx}(\mathbf{u}_3) = 12 > 0$, we conclude that $\mathbf{u}_1$ is a saddle point, $\mathbf{u}_2$ and $\mathbf{u}_3$ are local minimizers.

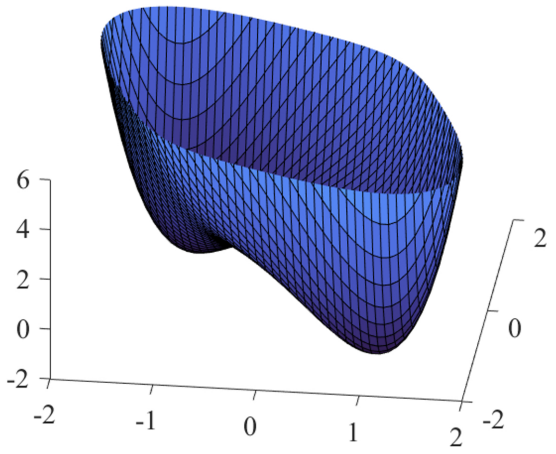


Figure 4.11: The function $f(x, y) = x^4 + y^4 + 4xy$.

Example 4.36

Consider the function $f : \mathbb{R}^3 \to \mathbb{R}$ defined as

$$f(x, y, z) = x^3 - xy^2 + 5x^2 - 4xy - 2xz + y^2 + 6yz + 37z^2.$$

Show that $(0, 0, 0)$ is a local minimizer of f.

Solution

Since f is a polynomial function, it is infinitely differentiable. Since

$$\nabla f(x, y, z) = (3x^2 - y^2 + 10x - 4y - 2z, -2xy - 4x + 2y + 6z, -2x + 6y + 74z),$$

we find that

$$\nabla f(0, 0, 0) = (0, 0, 0).$$

Hence, $(0, 0, 0)$ is a stationary point.
Now,

$$H_f(x, y, z) = \begin{bmatrix} 6x + 10 & -2y - 4 & -2 \\ -2y - 4 & -2x + 2 & 6 \\ -2 & 6 & 74 \end{bmatrix}.$$

Therefore,

$$H_f(0, 0, 0) = \begin{bmatrix} 10 & -4 & -2 \\ -4 & 2 & 6 \\ -2 & 6 & 74 \end{bmatrix}.$$

The determinants of the three principal submatrices of $H_f(0, 0, 0)$ are

$$\det M_1 = 10, \qquad \det M_2 = \begin{vmatrix} 10 & -4 \\ -4 & 2 \end{vmatrix} = 4,$$

$$\det M_3 = \begin{vmatrix} 10 & -4 & -2 \\ -4 & 2 & 6 \\ -2 & 6 & 74 \end{vmatrix} = 24.$$

This shows that $H_f(0, 0, 0)$ is positive definite. Hence, $(0, 0, 0)$ is a local minimizer of f.

Exercises 4.5

Question 1

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function defined as

$$f(x, y) = x^2 + 4y^2 + 5xy - 8x - 11y + 7.$$

Find the stationary points of f and classify them.

Question 2

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function defined as

$$f(x, y) = x^2 + 4y^2 + 3xy - 5x - 18y + 1.$$

Find the stationary points of f and classify them.

Question 3

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function defined as

$$f(x, y) = x^3 + y^3 + 12xy.$$

Find the stationary points of f and classify them.

Question 4

Consider the function $f : \mathbb{R}^3 \to \mathbb{R}$ defined as

$$f(x, y, z) = z^3 - 2z^2 - x^2 - y^2 - xy + x - y.$$

Show that $(1, -1, 0)$ is a stationary point of f and determine the nature of this stationary point.

Question 5

Consider the function $f : \mathbb{R}^3 \to \mathbb{R}$ defined as

$$f(x, y, z) = z^3 + 2z^2 - x^2 - y^2 - xy + x - y.$$

Show that $(1, -1, 0)$ is a stationary point of f and determine the nature of this stationary point.

Chapter 5

The Inverse and Implicit Function Theorems

In this chapter, we discuss the inverse function theorem and implicit function theorem, which are two important theorems in multivariable analysis. Given a function that maps a subset of $\mathbb{R}^n$ to $\mathbb{R}^n$, the inverse function theorem gives sufficient conditions for the existence of a local inverse and its differentiability. Given a system of m equations with $n+m$ variables, the implicit function theorem gives sufficient conditions to solve m of the variables in terms of the other n variables locally such that the solutions are differentiable functions. We want to emphasize that these theorems are *local*, in the sense that each of them asserts the existence of a function defined in a neighbourhood of a point.

In some sense, the two theorems are equivalent, which means one can deduce one from the other. In this book, we will prove the inverse function theorem first, and use it to deduce the implicit function theorem.

5.1 The Inverse Function Theorem

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$. If the function $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^n$ is one-to-one, we can define the inverse function $\mathbf{F}^{-1} : \mathbf{F}(\mathfrak{D}) \to \mathbb{R}^n$. The question we want to study here is the following. If $\mathfrak{D}$ is an open set and $\mathbf{F}$ is differentiable at the point $\mathbf{x}_0$ in $\mathfrak{D}$, is the inverse function $\mathbf{F}^{-1}$ differentiable at $\mathbf{y}_0 = \mathbf{F}(\mathbf{x}_0)$? For this, we also want the point $\mathbf{y}_0$ to be an interior point of $\mathbf{F}(\mathfrak{D})$. More precisely, is there a neighbourhood U of $\mathbf{x}_0$ that is mapped bijectively by $\mathbf{F}$ to a neighbourhood V of $\mathbf{y}_0$? If the answer is yes, and $\mathbf{F}^{-1}$ is differentiable at $\mathbf{y}_0$, then the chain rule would imply that

$$\mathbf{DF}^{-1}(\mathbf{y}_0)\mathbf{DF}(\mathbf{x}_0) = I_n.$$

Hence, a necessary condition for $\mathbf{F}^{-1}$ to be differentiable at $\mathbf{y}_0$ is that the derivative matrix $\mathrm{DF}(\mathbf{x}_0)$ has to be invertible.

Let us study the map $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^2$. The range of the function is $[0, \infty)$. Notice that if $x_0 > 0$, then $I = (0, \infty)$ is a neighbourhood of x_0 that is mapped bijectively by f to the neighbourhood $J = (0, \infty)$ of $f(x_0)$. If $x_0 < 0$, then $I = (-\infty, 0)$ is a neighbourhood of x_0 that is mapped bijectively by f to the neighbourhood $J = (0, \infty)$ of $f(x_0)$. However, if $x_0 = 0$, the point $f(x_0) = 0$ is not an interior point of $f(\mathbb{R}) = [0, \infty)$. Notice that $f'(x) = 2x$. Therefore, $x = 0$ is the point which $f'(x) = 0$.

If $x_0 > 0$, take $I = (0, \infty)$ and $J = (0, \infty)$. Then $f : I \to J$ has an inverse given by $f^{-1} : J \to I$, $f^{-1}(x) = \sqrt{x}$. It is a differentiable function with

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

In particular, at $y_0 = f(x_0) = x_0^2$,

$$(f^{-1})'(y_0) = \frac{1}{2\sqrt{y_0}} = \frac{1}{2x_0} = \frac{1}{f'(x_0)}.$$

Similarly, if $x_0 < 0$, take $I = (-\infty, 0)$ and $J = (0, \infty)$. Then $f : I \to J$ has an inverse given by $f^{-1} : J \to I$, $f^{-1}(x) = -\sqrt{x}$. It is a differentiable function with

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

In particular, at $y_0 = f(x_0) = x_0^2$,

$$(f^{-1})'(y_0) = -\frac{1}{2\sqrt{y_0}} = \frac{1}{2x_0} = \frac{1}{f'(x_0)}.$$

For a single variable function, the inverse function theorem takes the following form.

Theorem 5.1 (Single Variable) Inverse Function Theorem

Let $\mathcal{O}$ be an open subset of $\mathbb{R}$ that contains the point x_0, and let $f : \mathcal{O} \to \mathbb{R}$ be a continuously differentiable function defined on $\mathcal{O}$. Suppose that $f'(x_0) \neq 0$. Then there exists an open interval I containing x_0 such that f maps I bijectively onto the open interval $J = f(I)$. The inverse function $f^{-1} : J \to I$ is continuously differentiable. For any $y \in J$, if x is the point in I such that $f(x) = y$, then

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

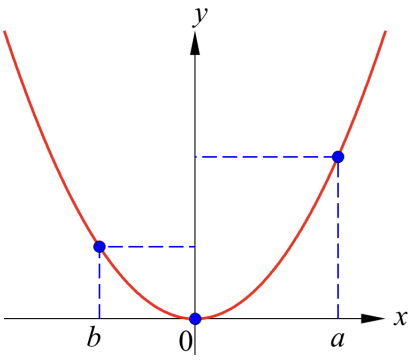


Figure 5.1: The function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$.

Proof

Without loss of generality, assume that $f'(x_0) > 0$. Since $\mathcal{O}$ is an open set and f' is continuous at x_0, there is an $r_1 > 0$ such that $(x_0 - r_1, x_0 + r_1) \subset \mathcal{O}$ and for all $x \in (x_0 - r_1, x_0 + r_1)$,

$$|f'(x) - f'(x_0)| < \frac{f'(x_0)}{2}.$$

This implies that

$$f'(x) > \frac{f'(x_0)}{2} > 0 \qquad \text{for all } x \in (x_0 - r_1, x_0 + r_1).$$

Therefore, f is strictly increasing on $(x_0 - r_1, x_0 + r_1)$. Take any $r > 0$ that is less that r_1. Then $[x - r, x + r] \subset (x_0 - r_1, x_0 + r_1)$. By intermediate value theorem, the function f maps $[x - r, x + r]$ bijectively onto $[f(x - r), f(x + r)]$. Let $I = (x - r, x + r)$ and $J = (f(x - r), f(x + r))$. Then $f : I \to J$ is a bijection and $f^{-1} : J \to I$ exists. In volume I, we have proved that f^{-1} is differentiable, and

$$(f^{-1})'(y) = \frac{1}{f'(f^{-1}(y))} \qquad \text{for all } y \in J.$$

This formula shows that $(f^{-1})' : J \to \mathbb{R}$ is continuous.

Remark 5.1

In the inverse function theorem, we determine the invertibility of the function in a neighbourhood of a point x_0. The theorem says that if f is continuously differentiable and $f'(x_0) \neq 0$, then f is locally invertible at x_0. Here the assumption that f' is continuous is essential. In volume I, we have seen that for a continuous function $f : I \to \mathbb{R}$ defined on an open interval I to be one-to-one, it is necessary that it is strictly monotonic. The function $f : \mathbb{R} \to \mathbb{R}$,

$$f(x) = \begin{cases} x + x^2 \sin\left(\dfrac{1}{x}\right), & \text{if } x \neq 0, \\ 0, & \text{if } x = 0, \end{cases}$$

is an example of a differentiable function where $f'(0) = 1 \neq 0$, but f fails to be strictly monotonic in any neighbourhood of the point $x = 0$.
This annoying behavior can be removed if we assume that f' is continuous. If $f'(x_0) \neq 0$ and f' is continuous, there is a neighbourhood I of x_0 such that $f'(x)$ has the same sign as $f'(x_0)$ for all $x \in I$. This implies that f is strictly monotonic on I.

Example 5.1

Let $f : \mathbb{R} \to \mathbb{R}$ be the function defined as

$$f(x) = 2x + 4\cos x.$$

Show that there is an open interval I containing 0 such that $f : I \to \mathbb{R}$ is one-to-one, and $f^{-1} : f(I) \to \mathbb{R}$ is continuously differentiable. Determine $(f^{-1})'(f(0))$.

Solution

The function f is infinitely differentiable and $f'(x) = 2 - 4\sin x$. Since $f'(0) = 2 \neq 0$, the inverse function theorem says that there is an open interval I containing 0 such that $f : I \to \mathbb{R}$ is one-to-one, and $f^{-1} : f(I) \to \mathbb{R}$ is continuously differentiable. Moreover,

$$(f^{-1})'(f(0)) = \frac{1}{f'(0)} = \frac{1}{2}.$$

Now let us consider functions defined on open subsets of $\mathbb{R}^n$, where $n \geq 2$. We first consider a linear transformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$. There is an $n \times n$ matrix A such that

$$\mathbf{T}(\mathbf{x}) = A\mathbf{x}.$$

The mapping $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$ is one-to-one if and only if A is invertible, if and only if $\det A \neq 0$. In this case, $\mathbf{T}$ is a bijection and $\mathbf{T}^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ is the linear transformation given by

$$\mathbf{T}^{-1}(\mathbf{x}) = A^{-1}\mathbf{x}.$$

Notice that for any $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$,

$$\mathbf{DT}(\mathbf{x}) = A, \qquad \mathbf{DT}^{-1}(\mathbf{y}) = A^{-1}.$$

The content of the inverse function theorem is to extend this to nonlinear mappings.

Theorem 5.2 Inverse Function Theorem

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^n$ be a continuously differentiable function defined on $\mathcal{O}$. If $\det \mathbf{DF}(\mathbf{x}_0) \neq 0$, then we have the followings.

(i) There exists a neighbourhood U of $\mathbf{x}_0$ such that $\mathbf{F}$ maps U bijectively onto the *open* set $V = \mathbf{F}(U)$.

(ii) The inverse function $\mathbf{F}^{-1} : V \to U$ is continuously differentiable.

(iii) For any $\mathbf{y} \in V$, if $\mathbf{x}$ is the point in U such that $\mathbf{F}(\mathbf{x}) = \mathbf{y}$, then

$$\mathbf{DF}^{-1}(\mathbf{y}) = \mathbf{DF}(\mathbf{F}^{-1}(\mathbf{y}))^{-1} = \mathbf{DF}(\mathbf{x})^{-1}.$$

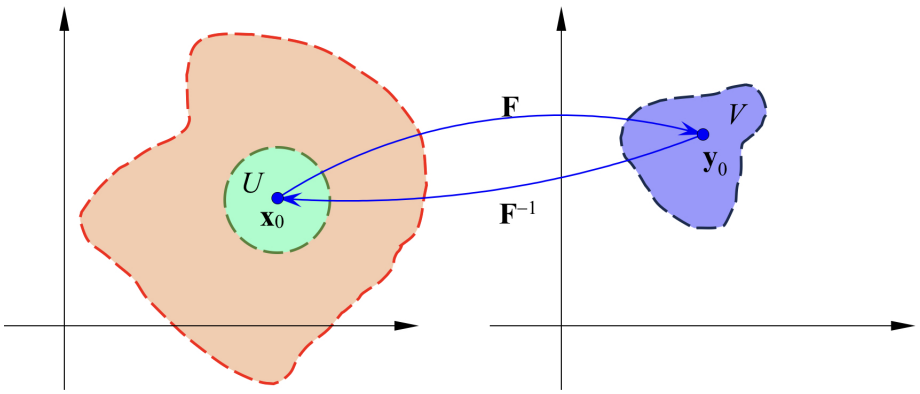


Figure 5.2: The inverse function theorem.

For a linear transformation which is a degree one polynomial mapping, the inverse function theorem holds *globally*. For a general continuously differentiable mapping, the inverse function theorem says that the first order approximation of the function at a point can determine the local invertibility of the function at that point.

When $n \geq 2$, the proof of the inverse function theorem is substantially more complicated than the $n = 1$ case, as we do not have the monotonicity argument used in the $n = 1$ case. The proof will be presented in Section 5.2. We will discuss the examples and applications in this section.

Example 5.2

Let $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$ be the mapping defined by

$$\mathbf{F}(x, y) = (3x - 2y + 7, 4x + 5y - 2).$$

Show that $\mathbf{F}$ is a bijection, and find $\mathbf{F}^{-1}(x, y)$ and $\mathbf{DF}^{-1}(x, y)$.

Solution

The mapping $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$ can be written as $\mathbf{F}(\mathbf{x}) = \mathbf{T}(\mathbf{x}) + \mathbf{b}$, where $\mathbf{T} : \mathbb{R}^2 \to \mathbb{R}^2$ is the linear transformation

$$\mathbf{T}(x, y) = (3x - 2y, 4x + 5y),$$

and $\mathbf{b} = (7, -2)$. For $\mathbf{u} = (x, y)$, $\mathbf{T}(\mathbf{u}) = A\mathbf{u}$, where $A = \begin{bmatrix} 3 & -2 \\ 4 & 5 \end{bmatrix}$. Since $\det A = 23 \neq 0$, the linear transformation $\mathbf{T} : \mathbb{R}^2 \to \mathbb{R}^2$ is one-to-one. Hence, $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$ is also one-to-one. Given $\mathbf{v} \in \mathbb{R}^2$, let $\mathbf{u} = A^{-1}(\mathbf{v} - \mathbf{b})$. Then $\mathbf{F}(\mathbf{u}) = \mathbf{v}$. Hence, $\mathbf{F}$ is also onto. The inverse $\mathbf{F}^{-1} : \mathbb{R}^2 \to \mathbb{R}^2$ is given by

$$\mathbf{F}^{-1}(\mathbf{v}) = A^{-1}(\mathbf{v} - \mathbf{b}).$$

Since

$$A^{-1} = \frac{1}{23}\begin{bmatrix} 5 & 2 \\ -4 & 3 \end{bmatrix},$$

we find that

$$\begin{aligned} \mathbf{F}^{-1}(x, y) &= \left(\frac{5(x-7) + 2(y+2)}{23}, \frac{-4(x-7) + 3(y+2)}{23} \right) \\ &= \left(\frac{5x + 2y - 31}{23}, \frac{-4x + 3y + 34}{23} \right), \end{aligned}$$

and

$$\mathbf{DF}^{-1}(x, y) = \frac{1}{23}\begin{bmatrix} 5 & 2 \\ -4 & 3 \end{bmatrix}.$$

Example 5.3

Determine the values of a such that the mapping $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^3$ defined by

$$\mathbf{F}(x, y, z) = (2x + y + az, x - y + 3z, 3x + 2y + z + 7)$$

is invertible.

Solution

The mapping $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^3$ can be written as $\mathbf{F}(\mathbf{x}) = \mathbf{T}(\mathbf{x}) + \mathbf{b}$, where $\mathbf{T} : \mathbb{R}^3 \to \mathbb{R}^3$ is the linear transformation

$$\mathbf{T}(x, y, z) = (2x + y + az, x - y + 3z, 3x + 2y + z),$$

and $\mathbf{b} = (0, 0, 7)$. Thus, $\mathbf{F}$ is a degree one polynomial mapping with

$$\mathbf{DF}(\mathbf{x}) = \begin{bmatrix} 2 & 1 & a \\ 1 & -1 & 3 \\ 3 & 2 & 1 \end{bmatrix}.$$

The mapping $\mathbf{F}$ is invertible if and only if it is one-to-one, if and only if $\mathbf{T}$ is one-to-one, if and only if $\det \mathbf{DF}(\mathbf{x}) \neq 0$. Since

$$\det \mathbf{DF}(\mathbf{x}) = 5a - 6,$$

the mapping $\mathbf{F}$ is invertible if and only if $a \neq 6/5$.

Example 5.4

Let $\mathbf{\Phi} : \mathbb{R}^2 \to \mathbb{R}^2$ be the mapping defined as

$$\mathbf{\Phi}(r, \theta) = (r \cos\theta, r \sin\theta).$$

Determine the points $(r, \theta) \in \mathbb{R}^2$ where the inverse function theorem can be applied to this mapping. Explain the significance of this result.

Solution

Since $\sin\theta$ and $\cos\theta$ are infinitely differentiable functions, the mapping $\mathbf{\Phi}$ is infinitely differentiable with

$$\mathbf{D\Phi}(r, \theta) = \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix}.$$

Since

$$\det \mathbf{D\Phi}(r, \theta) = r\cos^2\theta + r\sin^2\theta = r,$$

the inverse function theorem is not applicable at the point (r, θ) if $r = 0$. The mapping $\mathbf{\Phi}$ is a change from polar coordinates to rectangular coordinates. The result above shows that the change of coordinates is locally one-to-one away from the origin of the xy-plane.

Example 5.5

Consider the mapping $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$\mathbf{F}(x, y) = (x^2 - y^2, xy).$$

Show that there is a neighbourhood U of the point $\mathbf{u}_0 = (1, 1)$ such that $\mathbf{F} : U \to \mathbb{R}^2$ is one-to-one, $V = \mathbf{F}(U)$ is an open set, and $\mathbf{G} = \mathbf{F}^{-1} : V \to U$ is continuously differentiable. Then find $\dfrac{\partial G_1}{\partial y}(0, 1)$.

Solution

The mapping $\mathbf{F}$ is a polynomial mapping. Thus, it is continuously differentiable. Notice that $\mathbf{F}(\mathbf{u}_0) = (0, 1)$ and

$$\mathbf{DF}(x, y) = \begin{bmatrix} 2x & -2y \\ y & x \end{bmatrix}, \quad \mathbf{DF}(\mathbf{u}_0) = \begin{bmatrix} 2 & -2 \\ 1 & 1 \end{bmatrix}.$$

Since $\det \mathbf{DF}(\mathbf{u}_0) = 4 \neq 0$, the inverse function theorem implies that there is a neighbourhood U of the point $\mathbf{u}_0$ such that $\mathbf{F} : U \to \mathbb{R}^2$ is one-to-one, $V = \mathbf{F}(U)$ is an open set, and $\mathbf{G} = \mathbf{F}^{-1} : V \to U$ is continuously differentiable. Moreover,

$$\mathbf{DG}(0, 1) = \mathbf{DF}(1, 1)^{-1} = \frac{1}{4} \begin{bmatrix} 1 & 2 \\ -1 & 2 \end{bmatrix}.$$

From here, we find that

$$\frac{\partial G_1}{\partial y}(0, 1) = \frac{2}{4} = \frac{1}{2}.$$

Example 5.6

Consider the system of equations

$$\begin{aligned} \sin(x + y) + x^2 y + 3xy^2 &= 2, \\ 2xy + 5x^2 - 2y^2 &= 1. \end{aligned}$$

Observe that $(x, y) = (1, -1)$ is a solution of this system. Show that there is a neighbourhood U of $\mathbf{u}_0 = (1, -1)$ and an $r > 0$ such that for all (a, b) satisfying $(a-2)^2 + (b-1)^2 < r^2$, the system

$$\begin{aligned} \sin(x+y) + x^2 y + 3xy^2 &= a, \\ 2xy + 5x^2 - 2y^2 &= b \end{aligned}$$

has a unique solution (x, y) that lies in U.

Solution

Let $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$ be the function defined by

$$\mathbf{F}(x, y) = \left(\sin(x+y) + x^2 y + 3xy^2, 2xy + 5x^2 - 2y^2\right).$$

Since the sine function is infinitely differentiable, $\sin(x+y)$ is infinitely differentiable. The functions $g(x, y) = x^2 y + 3xy^2$ and $F_2(x, y) = 2xy + 5x^2 - 2y^2$ are polynomial functions. Hence, they are also infinitely differentiable. This shows that $\mathbf{F}$ is infinitely differentiable. Since

$$\mathbf{DF}(x, y) = \begin{bmatrix} \cos(x+y) + 2xy + 3y^2 & \cos(x+y) + x^2 + 6xy \\ 2y + 10x & 2x - 4y \end{bmatrix},$$

we find that

$$\mathbf{DF}(1, -1) = \begin{bmatrix} 2 & -4 \\ 8 & 6 \end{bmatrix}.$$

It follows that $\det \mathrm{DF}(1, -1) = 44 \neq 0$.

By the inverse function theorem, there exists a neighbourhood U_1 of $\mathbf{u}_0$ such that $\mathbf{F} : U_1 \to \mathbb{R}^2$ is one-to-one and $V = \mathbf{F}(U_1)$ is an open set. Since $\mathbf{F}(\mathbf{u}_0) = (2, 1)$, the point $\mathbf{v}_0 = (2, 1)$ is a point in the open set V. Hence, there exists $r > 0$ such that $B(\mathbf{v}_0, r) \subset V$. Since $B(\mathbf{v}_0, r)$ is open and $\mathbf{F}$ is continuous, $U = \mathbf{F}^{-1}\left(B(\mathbf{v}_0, r)\right)$ is an open subset of $\mathbb{R}^2$. The map $\mathbf{F} : U \to B(\mathbf{v}_0, r)$ is a bijection. For all (a, b) satisfying $(a-2)^2 + (b-1)^2 < r^2$, (a, b) is in $B(\mathbf{v}_0, r)$. Hence, there is a unique (x, y) in U such that $\mathbf{F}(x, y) = (a, b)$. This means that the system

$$\sin(x+y)+x^2y+3xy^2=a,$$
$$2xy+5x^2-2y^2=b$$

has a unique solution (x,y) that lies in U.

At the end of this section, let us prove the following theorem.

Theorem 5.3

Let A be an $n\times n$ matrix, and let $\mathbf{x}_0$ and $\mathbf{y}_0$ be two points in $\mathbb{R}^n$. Define the mapping $\mathbf{F}:\mathbb{R}^n\to\mathbb{R}^n$ by

$$\mathbf{F}(\mathbf{x})=\mathbf{y}_0+A\left(\mathbf{x}-\mathbf{x}_0\right).$$

Then $\mathbf{F}$ is infinitely differentiable with $\mathbf{DF}(\mathbf{x})=A$. It is one-to-one and onto if and only if $\det A\neq 0$. In this case,

$$\mathbf{F}^{-1}(\mathbf{y})=\mathbf{x}_0+A^{-1}\left(\mathbf{y}-\mathbf{y}_0\right),\quad \text{and}\quad \mathbf{DF}^{-1}(\mathbf{y})=A^{-1}.$$

In particular, $\mathbf{F}^{-1}$ is also infinitely differentiable.

Proof

Obviously, $\mathbf{F}$ is a polynomial mapping. Hence, $\mathbf{F}$ is infinitely differentiable. By a straightforward computation, we find that $\mathbf{DF}=A$.

Notice that $\mathbf{F}=\mathbf{F}_2\circ\mathbf{T}\circ\mathbf{F}_1$, where $\mathbf{F}_1:\mathbb{R}^n\to\mathbb{R}^n$ is the translation $\mathbf{F}_1(\mathbf{x})=\mathbf{x}-\mathbf{x}_0$, $\mathbf{T}:\mathbb{R}^n\to\mathbb{R}^n$ is the linear transformation $\mathbf{T}(\mathbf{x})=A\mathbf{x}$, and $\mathbf{F}_2:\mathbb{R}^n\to\mathbb{R}^n$ is the translation $\mathbf{F}_2(\mathbf{y})=\mathbf{y}+\mathbf{y}_0$. Since translations are bijective mappings, $\mathbf{F}$ is one-to-one and onto if and only if $\mathbf{T}:\mathbb{R}^n\to\mathbb{R}^n$ is one-to-one and onto, if and only if $\det A\neq 0$.

If

$$\mathbf{y}=\mathbf{y}_0+A\left(\mathbf{x}-\mathbf{x}_0\right),$$

then

$$\mathbf{x}=\mathbf{x}_0+A^{-1}\left(\mathbf{y}-\mathbf{y}_0\right).$$

This gives the formula for $\mathbf{F}^{-1}(\mathbf{y})$. The formula for $\mathbf{DF}^{-1}(\mathbf{y})$ follows.

Exercises 5.1

Question 1

Let $f : \mathbb{R} \to \mathbb{R}$ be the function defined as

$$f(x) = e^{2x} + 4x \sin x + 2 \cos x.$$

Show that there is an open interval I containing 0 such that $f : I \to \mathbb{R}$ is one-to-one, and $f^{-1} : f(I) \to \mathbb{R}$ is continuously differentiable. Determine $(f^{-1})'(f(0))$.

Question 2

Let $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$ be the mapping defined by

$$\mathbf{F}(x, y) = (3x + 2y - 5, 7x + 4y - 3).$$

Show that $\mathbf{F}$ is a bijection, and find $\mathbf{F}^{-1}(x, y)$ and $\mathbf{DF}^{-1}(x, y)$.

Question 3

Consider the mapping $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$\mathbf{F}(x, y) = (x^2 + y^2, xy).$$

Show that there is a neighbourhood U of the point $\mathbf{u}_0 = (2, 1)$ such that $\mathbf{F} : U \to \mathbb{R}^2$ is one-to-one, $V = \mathbf{F}(U)$ is an open set, and $\mathbf{G} = \mathbf{F}^{-1} : V \to U$ is continuously differentiable. Then find $\dfrac{\partial G_2}{\partial x}(5, 2)$.

Question 4

Let $\mathbf{\Phi} : \mathbb{R}^3 \to \mathbb{R}^3$ be the mapping defined as

$$\mathbf{\Phi}(\rho, \phi, \theta) = (\rho \sin \phi \cos \theta, \rho \sin \phi \sin \theta, \rho \cos \phi).$$

Determine the points $(\rho, \phi, \theta) \in \mathbb{R}^3$ where the inverse function theorem can be applied to this mapping. Explain the significance of this result.

Question 5

Consider the system of equations

$$\begin{aligned} 4x + y - 5xy &= 2, \\ x^2 + y^2 - 3xy^2 &= 5. \end{aligned}$$

Observe that $(x, y) = (-1, 1)$ is a solution of this system. Show that there is a neighbourhood U of $\mathbf{u}_0 = (-1, 1)$ and an $r > 0$ such that for all (a, b) satisfying $(a-2)^2 + (b-5)^2 < r^2$, the system

$$\begin{aligned} 4x + y - 5xy &= a, \\ x^2 + y^2 - 3xy^2 &= b \end{aligned}$$

has a unique solution (x, y) that lies in U.

5.2 The Proof of the Inverse Function Theorem

In this section, we prove the inverse function theorem stated in Theorem 5.2. The hardest part of the proof is the first statement, which asserts that there is a neighbourhood U of $\mathbf{x}_0$ such that restricted to U, $\mathbf{F}$ is one-to-one, and the image of U under $\mathbf{F}$ is open in $\mathbb{R}^n$.

In the statement of the inverse function theorem, we assume that the derivative matrix of the continuously differentiable mapping $\mathbf{F} : \mathcal{O} \to \mathbb{R}^n$ is invertible at the point $\mathbf{x}_0$. The continuities of the partial derivatives of $\mathbf{F}$ then implies that there is a neighbourhood $\mathcal{N}$ of $\mathbf{x}_0$ such that the derivative matrix of $\mathbf{F}$ at any $\mathbf{x}$ in $\mathcal{N}$ is also invertible.

Theorem 3.38 asserts that a linear transformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$ is invertible if and only if there is a positive constant c such that

$$\|\mathbf{T}(\mathbf{u}) - \mathbf{T}(\mathbf{v})\| \geq c\|\mathbf{u} - \mathbf{v}\| \qquad \text{for all } \mathbf{u}, \mathbf{v} \in \mathbb{R}^n.$$

Definition 5.1 Stable Mappings

A mapping $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^n$ is *stable* if there is a positive constant c such that

$$\|\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v})\| \geq c\|\mathbf{u} - \mathbf{v}\| \qquad \text{for all } \mathbf{u}, \mathbf{v} \in \mathfrak{D}.$$

In other words, a linear transformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$ is invertible if and only if it is stable.

Remark 5.2 Stable Mappings vs Lipschitz Mappings

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$. Observe that if $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^n$ is a stable mapping, there is a constant $c > 0$ such that

$$\|\mathbf{F}(\mathbf{u}_1) - \mathbf{F}(\mathbf{u}_2)\| \geq c\|\mathbf{u}_1 - \mathbf{u}_2\| \qquad \text{for all } \mathbf{u}_1, \mathbf{u}_2 \in \mathfrak{D}.$$

This implies that $\mathbf{F}$ is one-to-one, and thus the inverse $\mathbf{F}^{-1} : \mathbf{F}(\mathfrak{D}) \to \mathbb{R}^n$ exists. Notice that for any $\mathbf{v}_1$ and $\mathbf{v}_2$ in $\mathbf{F}(\mathfrak{D})$,

$$\|\mathbf{F}^{-1}(\mathbf{v}_1) - \mathbf{F}^{-1}(\mathbf{v}_2)\| \leq \frac{1}{c}\|\mathbf{v}_1 - \mathbf{v}_2\|.$$

This means that $\mathbf{F}^{-1} : \mathbf{F}(\mathfrak{D}) \to \mathbb{R}^n$ is a Lipschitz mapping.

For a mapping $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^n$ that satisfies the assumptions in the statement of the inverse function theorem, it is stable in a neighbourhood of $\mathbf{x}_0$.

Theorem 5.4

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^n$ be a continuously differentiable function defined on $\mathcal{O}$. If $\det \mathbf{DF}(\mathbf{x}_0) \neq 0$, then there exists a neighbourhood U of $\mathbf{x}_0$ such that $\mathbf{DF}(\mathbf{x})$ is invertible for all $\mathbf{x} \in U$, $\mathbf{F}$ maps U bijectively onto the open set $V = \mathbf{F}(U)$, and the map $\mathbf{F} : U \to V$ is stable.

Recall that when A is a subset of $\mathbb{R}^n$, $\mathbf{u}$ is a point in $\mathbb{R}^n$,

$$A + \mathbf{u} = \{\mathbf{a} + \mathbf{u} \,|\, \mathbf{a} \in A\}$$

is the translate of the set A by the vector $\mathbf{u}$. The set A is open if and only if $A + \mathbf{u}$ is open, A is closed if and only if $A + \mathbf{u}$ is closed.

Lemma 5.5

It is sufficient to prove Theorem 5.4 when $\mathbf{x}_0 = \mathbf{0}$, $\mathbf{F}(\mathbf{x}_0) = \mathbf{0}$ and $\mathbf{DF}(\mathbf{x}_0) = I_n$.

Proof of Lemma 5.5

Assume that Theorem 5.4 holds when $\mathbf{x}_0 = \mathbf{0}$, $\mathbf{F}(\mathbf{x}_0) = \mathbf{0}$ and $\mathbf{DF}(\mathbf{x}_0) = I_n$.

Now given that $\mathbf{F} : \mathcal{O} \to \mathbb{R}^n$ is a continuously differentiable mapping with $\det \mathbf{DF}(\mathbf{x}_0) \neq 0$, let $\mathbf{y}_0 = \mathbf{F}(\mathbf{x}_0)$ and $A = \mathbf{DF}(\mathbf{x}_0)$. Then A is invertible. Define the open set $\mathcal{D}$ as $\mathcal{D} = \mathcal{O} - \mathbf{x}_0$. It is a neighbourhood of the point $\mathbf{0}$. Let $\mathbf{G} : \mathcal{D} \to \mathbb{R}^n$ be the mapping

$$\mathbf{G}(\mathbf{x}) = A^{-1} \left(\mathbf{F}(\mathbf{x} + \mathbf{x}_0) - \mathbf{y}_0\right).$$

Then $\mathbf{G}(\mathbf{0}) = \mathbf{0}$. Using the same reasoning as the proof of Theorem 5.3, we find that $\mathbf{G}$ is continuously differentiable and

$$\mathbf{DG}(\mathbf{x}) = A^{-1}\mathbf{DF}(\mathbf{x} + \mathbf{x}_0).$$

This gives

$$\mathbf{DG}(\mathbf{0}) = A^{-1}\mathbf{DF}(\mathbf{x}_0) = I_n.$$

By assumption, Theorem 5.4 holds for the mapping $\mathbf{G}$. Namely, there exist neighbourhoods $\mathcal{U}$ and $\mathcal{V}$ of $\mathbf{0}$ such that $\mathbf{G} : \mathcal{U} \to \mathcal{V}$ is a bijection and $\mathbf{DG}(\mathbf{x})$ is invertible for all $\mathbf{x} \in \mathcal{U}$. Moreover, there is a positive constant a such that

$$\|\mathbf{G}(\mathbf{u}_1) - \mathbf{G}(\mathbf{u}_2)\| \geq a\|\mathbf{u}_1 - \mathbf{u}_2\| \qquad \text{for all } \mathbf{u}_1, \mathbf{u}_2 \in \mathcal{U}.$$

Let U be the neighbourhood of $\mathbf{x}_0$ given by $U = \mathcal{U} + \mathbf{x}_0$. By Theorem 5.3, the mapping $\mathbf{H} : \mathbb{R}^n \to \mathbb{R}^n$,

$$\mathbf{H}(\mathbf{y}) = A^{-1}(\mathbf{y} - \mathbf{y}_0)$$

is a continuous bijection. Therefore, $V = \mathbf{H}^{-1}(\mathcal{V})$ is an open subset of $\mathbb{R}^n$ that contains $\mathbf{y}_0$. By definition, $\mathbf{F}$ maps U bijectively to V. Since

$$\mathbf{F}(\mathbf{x}) = \mathbf{y}_0 + A\mathbf{G}(\mathbf{x} - \mathbf{x}_0),$$

we find that

$$\mathbf{DF}(\mathbf{x}) = A\left(\mathbf{DG}(\mathbf{x} - \mathbf{x}_0)\right).$$

Since A is invertible, $\mathbf{DF}(\mathbf{x})$ is invertible for all $\mathbf{x} \in U$. Theorem 3.38 says that there is a positive constant α such that

$$\|A\mathbf{x}\| \geq \alpha\|\mathbf{x}\| \qquad \text{for all } \mathbf{x} \in \mathbb{R}^n.$$

Therefore, for any $\mathbf{u}_1$ and $\mathbf{u}_2$ in U,

$$\begin{aligned}\|\mathbf{F}(\mathbf{u}_1) - \mathbf{F}(\mathbf{u}_2)\| &= \|A\left(\mathbf{G}(\mathbf{u}_1 - \mathbf{x}_0) - \mathbf{G}(\mathbf{u}_2 - \mathbf{x}_0)\right)\| \\ &\geq \alpha\|\mathbf{G}(\mathbf{u}_1 - \mathbf{x}_0) - \mathbf{G}(\mathbf{u}_2 - \mathbf{x}_0)\| \\ &\geq a\alpha\|\mathbf{u}_1 - \mathbf{u}_2\|.\end{aligned}$$

This shows that $\mathbf{F} : U \to V$ is stable, and thus completes the proof of the lemma.

Now we prove Theorem 5.4.

Proof of Theorem 5.4

By Lemma 5.5, we only need to consider the case where $\mathbf{x}_0 = \mathbf{0}$, $\mathbf{F}(\mathbf{x}_0) = \mathbf{0}$ and $\mathbf{DF}(\mathbf{x}_0) = I_n$.

Since $\mathbf{F} : \mathcal{O} \to \mathbb{R}^n$ is continuously differentiable, the map $\mathbf{DF} : \mathcal{O} \to \mathcal{M}_n$ is continuous. Since $\det : \mathcal{M}_n \to \mathbb{R}$ is also continuous, and $\det \mathbf{DF}(\mathbf{0}) = 1$, there is an $r_0 > 0$ such that $B(\mathbf{0}, r_0) \subset \mathcal{O}$ and for all $\mathbf{x} \in B(\mathbf{0}, r_0)$, $\det \mathbf{DF}(\mathbf{x}) > \frac{1}{2}$. In particular, $\mathbf{DF}(\mathbf{x})$ is invertible for all $\mathbf{x} \in B(\mathbf{0}, r_0)$.

Let $\mathbf{G} : \mathcal{O} \to \mathbb{R}^n$ be the mapping defined as

$$\mathbf{G}(\mathbf{x}) = \mathbf{F}(\mathbf{x}) - \mathbf{x},$$

so that $\mathbf{F}(\mathbf{x}) = \mathbf{x} + \mathbf{G}(\mathbf{x})$. The mapping $\mathbf{G}$ is continuosly differentiable. It satisfies $\mathbf{G}(\mathbf{0}) = \mathbf{0}$ and

$$\mathbf{DG}(\mathbf{0}) = \mathbf{DF}(\mathbf{0}) - I_n = \mathbf{0}.$$

Since $\mathbf{G}$ is continuously differentiable, for any $1 \le i \le n$, $1 \le j \le n$, there exists $r_{i,j} > 0$ such that $B(\mathbf{0}, r_{i,j}) \subset \mathcal{O}$ and for all $\mathbf{x} \in B(\mathbf{0}, r_{i,j})$,

$$\left|\frac{\partial G_i}{\partial x_j}(\mathbf{x})\right| = \left|\frac{\partial G_i}{\partial x_j}(\mathbf{x}) - \frac{\partial G_i}{\partial x_j}(\mathbf{0})\right| < \frac{1}{2n}.$$

Let

$$r = \min\left(\{r_{i,j} \,|\, 1 \le i \le n, 1 \le j \le n\} \cup \{r_0\}\right).$$

Then $r > 0$, $B(\mathbf{0}, r) \subset B(\mathbf{0}, r_0)$ and $B(\mathbf{0}, r) \subset B(\mathbf{0}, r_{i,j})$ for all $1 \le i \le n$, $1 \le j \le n$. The ball $B(\mathbf{0}, r)$ is a convex set. If $\mathbf{u}$ and $\mathbf{v}$ are two points in $B(\mathbf{0}, r)$, mean value theorem implies that for $1 \le i \le n$, there exists $\mathbf{z}_i \in B(\mathbf{0}, r)$ such that

$$G_i(\mathbf{u}) - G_i(\mathbf{v}) = \sum_{j=1}^{n} (u_j - v_j) \frac{\partial G_i}{\partial x_j}(\mathbf{z}_i).$$

It follows that

$$\begin{aligned} |G_i(\mathbf{u}) - G_i(\mathbf{v})| &\le \sum_{j=1}^{n} |u_j - v_j| \left|\frac{\partial G_i}{\partial x_j}(\mathbf{z}_i)\right| \\ &\le \frac{1}{2n} \sum_{j=1}^{n} |u_j - v_j| \le \frac{1}{2\sqrt{n}} \|\mathbf{u} - \mathbf{v}\|. \end{aligned}$$

Therefore,

$$\|\mathbf{G}(\mathbf{u}) - \mathbf{G}(\mathbf{v})\| = \sqrt{\sum_{i=1}^{n} \left(G_i(\mathbf{u}) - G_i(\mathbf{v})\right)^2} \leq \frac{1}{2}\|\mathbf{u} - \mathbf{v}\|.$$

This shows that $\mathbf{G} : B(\mathbf{0}, r) \to \mathbb{R}^n$ is a map satisfying $\mathbf{G}(\mathbf{0}) = \mathbf{0}$, and

$$\|\mathbf{G}(\mathbf{u}) - \mathbf{G}(\mathbf{v})\| \leq \frac{1}{2}\|\mathbf{u} - \mathbf{v}\| \qquad \text{for all } \mathbf{u}, \mathbf{v} \in B(\mathbf{0}, r).$$

By Theorem 2.44, the map $\mathbf{F} : B(\mathbf{0}, r) \to \mathbb{R}^n$ is one-to-one, and its image contains the open ball $B(\mathbf{0}, r/2)$. Let $V = B(\mathbf{0}, r/2)$. Then V is an open subset of $\mathbb{R}^n$ that is contained in the image of $\mathbf{F}$. Since $\mathbf{F} : B(\mathbf{0}, r) \to \mathbb{R}^n$ is continuous, $U = \mathbf{F}|^{-1}_{B(\mathbf{0},r)}(V)$ is an open set. By definition, $\mathbf{F} : U \to V$ is a bijection. Since U is contained in $B(\mathbf{0}, r_0)$, $\mathbf{DF}(\mathbf{x})$ is invertible for all $\mathbf{x}$ in U. Finally, for any $\mathbf{u}$ and $\mathbf{v}$ in U,

$$\|\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v})\| \geq \|\mathbf{u} - \mathbf{v}\| - \|\mathbf{G}(\mathbf{u}) - \mathbf{G}(\mathbf{v})\| \geq \frac{1}{2}\|\mathbf{u} - \mathbf{v}\|.$$

This completes the proof of the theorem.

To complete the proof of the inverse function theorem, it remains to prove that $\mathbf{F}^{-1} : V \to U$ is continuously differentiable, and

$$\mathbf{DF}^{-1}(\mathbf{y}) = \mathbf{DF}(\mathbf{F}^{-1}(\mathbf{y}))^{-1}.$$

Theorem 5.6

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}_0$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^n$ be a continuously differentiable function defined on $\mathcal{O}$. If $\det \mathbf{DF}(\mathbf{x}_0) \neq 0$, then there exists a neighbourhood U of $\mathbf{x}_0$ such that $\mathbf{F}$ maps U bijectively onto the open set $V = \mathbf{F}(U)$, the inverse function $\mathbf{F}^{-1} : V \to U$ is continuously differentiable, and for any $\mathbf{y} \in V$, if $\mathbf{x}$ is the point in U such that $\mathbf{F}(\mathbf{x}) = \mathbf{y}$, then

$$\mathbf{DF}^{-1}(\mathbf{y}) = \mathbf{DF}(\mathbf{x})^{-1}.$$

Proof

Theorem 5.4 asserts that there exists a neighbourhood U of $\mathbf{x}_0$ such that $\mathbf{F}$ maps U bijectively onto the open set $V = \mathbf{F}(U)$, $\mathbf{DF}(\mathbf{x})$ is invertible for all $\mathbf{x}$ in U, and there is a positive constant c such that

$$\|\mathbf{F}(\mathbf{u}_1) - \mathbf{F}(\mathbf{u}_2)\| \geq c\|\mathbf{u}_1 - \mathbf{u}_2\| \qquad \text{for all } \mathbf{u}_1, \mathbf{u}_2 \in U. \tag{5.1}$$

Now given $\mathbf{y}$ in V, we want to show that $\mathbf{F}^{-1}$ is differentiable at $\mathbf{y}$ and $\mathbf{DF}^{-1}(\mathbf{y}) = \mathbf{DF}(\mathbf{x})^{-1}$, where $\mathbf{x} = \mathbf{F}^{-1}(\mathbf{y})$. Since V is open, there is an $r > 0$ such that $B(\mathbf{y}, r) \subset V$. For $\mathbf{k} \in \mathbb{R}^n$ such that $\|\mathbf{k}\| < r$, let

$$\mathbf{h}(\mathbf{k}) = \mathbf{F}^{-1}(\mathbf{y} + \mathbf{k}) - \mathbf{F}^{-1}(\mathbf{y}).$$

Then

$$\mathbf{F}(\mathbf{x}) = \mathbf{y} \quad \text{and} \quad \mathbf{F}(\mathbf{x} + \mathbf{h}) = \mathbf{y} + \mathbf{k}.$$

Eq. (5.1) implies that

$$\|\mathbf{h}\| \leq \frac{1}{c}\|\mathbf{k}\|. \tag{5.2}$$

Let $A = \mathbf{DF}(\mathbf{x})$. By assumption, A is invertible. Notice that

$$\begin{aligned}\mathbf{F}^{-1}(\mathbf{y} + \mathbf{k}) - \mathbf{F}^{-1}(\mathbf{y}) - A^{-1}\mathbf{k} &= -A^{-1}\left(\mathbf{k} - A\mathbf{h}\right) \\ &= -A^{-1}\left(\mathbf{F}(\mathbf{x} + \mathbf{h}) - \mathbf{F}(\mathbf{x}) - A\mathbf{h}\right).\end{aligned}$$

There is a positive constant β such that

$$\|A^{-1}\mathbf{y}\| \leq \beta\|\mathbf{y}\| \qquad \text{for all } \mathbf{y} \in \mathbb{R}^n.$$

Therefore,

$$\begin{aligned}&\left\|\frac{\mathbf{F}^{-1}(\mathbf{y} + \mathbf{k}) - \mathbf{F}^{-1}(\mathbf{y}) - A^{-1}\mathbf{k}}{\|\mathbf{k}\|}\right\| \\ &\leq \frac{\beta}{\|\mathbf{k}\|}\|\left(\mathbf{F}(\mathbf{x} + \mathbf{h}) - \mathbf{F}(\mathbf{x}) - A\mathbf{h}\right)\| \\ &\leq \frac{\beta}{c}\left\|\frac{\mathbf{F}(\mathbf{x} + \mathbf{h}) - \mathbf{F}(\mathbf{x}) - A\mathbf{h}}{\|\mathbf{h}\|}\right\|.\end{aligned} \tag{5.3}$$

Since $\mathbf{F}$ is differentiable at $\mathbf{x}$,

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\mathbf{F}(\mathbf{x} + \mathbf{h}) - \mathbf{F}(\mathbf{x}) - A\mathbf{h}}{\|\mathbf{h}\|} = \mathbf{0}.$$

Eq. (5.2) implies that $\lim_{\mathbf{k}\to\mathbf{0}} \mathbf{h} = \mathbf{0}$. Eq. (5.3) then implies that

$$\lim_{\mathbf{k}\to\mathbf{0}} \frac{\mathbf{F}^{-1}(\mathbf{y}+\mathbf{k}) - \mathbf{F}^{-1}(\mathbf{y}) - A^{-1}\mathbf{k}}{\|\mathbf{k}\|} = \mathbf{0}.$$

This proves that $\mathbf{F}^{-1}$ is differentiable at $\mathbf{y}$ and

$$\mathbf{DF}^{-1}(\mathbf{y}) = A^{-1} = \mathbf{DF}(\mathbf{x})^{-1}.$$

Now the map $\mathbf{DF}^{-1} : V \to \mathrm{GL}(n,\mathbb{R})$ is the compositions of the maps $\mathbf{F}^{-1} : V \to U$, $\mathbf{DF} : U \to \mathrm{GL}(n,\mathbb{R})$ and $\mathscr{I} : \mathrm{GL}(n,\mathbb{R}) \to \mathrm{GL}(n,\mathbb{R})$ which takes A to A^{-1}. Since each of these maps is continuous, the map $\mathbf{DF}^{-1} : V \to \mathrm{GL}(n,\mathbb{R})$ is continuous. This completes the proof that $\mathbf{F}^{-1} : V \to U$ is continuously differentiable.

At the end of this section, let us give a brief discussion about the concept of homeomorphism and diffeomorphism.

Definition 5.2 Homeomorphism

Let A be a subset of $\mathbb{R}^m$ and let B be a subset of $\mathbb{R}^n$. We say that A and B are homeomorphic if there exists a continuous bijective function $\mathbf{F} : A \to B$ whose inverse $\mathbf{F}^{-1} : B \to A$ is also continuous. Such a function $\mathbf{F}$ is called a homeomorphism between A and B.

Definition 5.3 Diffeomorphism

Let $\mathcal{O}$ and $\mathcal{U}$ be open subsets of $\mathbb{R}^n$. We say that $\mathcal{U}$ and $\mathcal{O}$ are diffeomorphic if there exists a homeomorphism $\mathbf{F} : \mathcal{O} \to \mathcal{U}$ between $\mathcal{O}$ and $\mathcal{U}$ such that $\mathbf{F}$ and $\mathbf{F}^{-1}$ are differentiable.

Example 5.7

Let $A = \{(x,y) \,|\, x^2 + y^2 < 1\}$ and $B = \{(x,y) \,|\, 4x^2 + 9y^2 < 36\}$. Define the map $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$ by

$$\mathbf{F}(x,y) = (3x, 2y).$$

Then $\mathbf{F}$ is an invertible linear transformation with

$$\mathbf{F}^{-1}(x, y) = \left(\frac{x}{3}, \frac{y}{2}\right).$$

The mappings $\mathbf{F}$ and $\mathbf{F}^{-1}$ are continuously differentiable. It is easy to show that $\mathbf{F}$ maps A bijectively onto B. Hence, $\mathbf{F} : A \to B$ is a diffeomorphism between A and B.

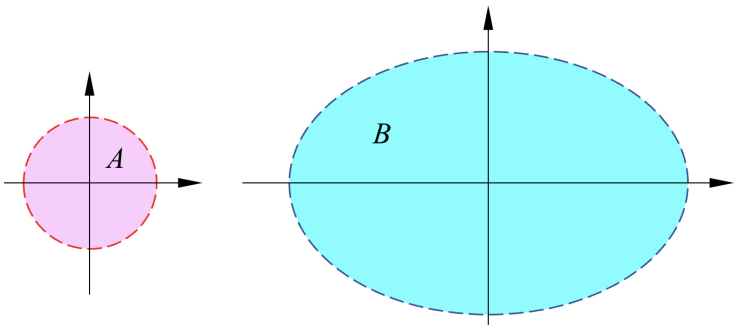


Figure 5.3: $A = \{(x, y) \,|\, x^2 + y^2 < 1\}$ and $B = \{(x, y) \,|, 4x^2 + 9y^2 < 36\}$ are diffeomorphic.

Theorem 5.3 gives the following.

Theorem 5.7

Let A be an invertible $n \times n$ matrix, and let $\mathbf{x}_0$ and $\mathbf{y}_0$ be two points in $\mathbb{R}^n$. Define the mapping $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^n$ by

$$\mathbf{F}(\mathbf{x}) = \mathbf{y}_0 + A\,(\mathbf{x} - \mathbf{x}_0)\,.$$

If $\mathcal{O}$ is an open subset of $\mathbb{R}^n$, then $\mathbf{F} : \mathcal{O} \to \mathbf{F}(\mathcal{O})$ is a diffeomorphism.

The inverse function theorem gives the following.

Theorem 5.8

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^n$ be a continuously differentiable mapping such that $\mathbf{DF}(\mathbf{x})$ is invertible for all $\mathbf{x} \in \mathcal{O}$. If $\mathcal{U}$ is an open subset contained in $\mathcal{O}$ such that $\mathbf{F} : \mathcal{U} \to \mathbb{R}^n$ is one-to-one, then $\mathbf{F} : \mathcal{U} \to \mathbf{F}(\mathcal{U})$ is a diffeomorphism.

The proof of this theorem is left as an exercise.

Exercises 5.2

Question 1

Let $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$ be the mapping given by

$$\mathbf{F}(x, y) = (xe^y + xy, 2x^2 + 3y^2).$$

Show that there is a neighbourhood U of $(-1, 0)$ such that the mapping $\mathbf{F} : U \to \mathbb{R}^2$ is stable.

Question 2

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^n$ be a continuously differentiable mapping such that $\det \mathbf{DF}(\mathbf{x}) \neq 0$ for all $\mathbf{x} \in \mathcal{O}$. Show that $\mathbf{F}(\mathcal{O})$ is an open set.

Question 3

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^n$ be a continuously differentiable mapping such that $\mathbf{DF}(\mathbf{x})$ is invertible for all $\mathbf{x} \in \mathcal{O}$. If $\mathcal{U}$ is an open subset contained in $\mathcal{O}$ such that $\mathbf{F} : \mathcal{U} \to \mathbb{R}^n$ is one-to-one, then $\mathbf{F} : \mathcal{U} \to \mathbf{F}(\mathcal{U})$ is a diffeomorphism.

Question 4

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^n$ be a differentiable mapping. Assume that there is a positive constant c such that

$$\|\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v})\| \geq c\|\mathbf{u} - \mathbf{v}\| \qquad \text{for all } \mathbf{u}, \mathbf{v} \in \mathcal{O}.$$

Use first order approximation theorem to show that for any $\mathbf{x} \in \mathcal{O}$ and any $\mathbf{h} \in \mathbb{R}^n$,

$$\|\mathbf{DF}(\mathbf{x})\mathbf{h}\| \geq c\|\mathbf{h}\|.$$

Question 5

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^n$ be a continuously differentiable mapping.

(a) If $\mathbf{F} : \mathcal{O} \to \mathbb{R}^n$ is stable, show that the derivative matrix $\mathbf{DF}(\mathbf{x})$ is invertible at every $\mathbf{x}$ in $\mathcal{O}$.

(b) Assume that the derivative matrix $\mathbf{DF}(\mathbf{x})$ is invertible at every $\mathbf{x}$ in $\mathcal{O}$. If C is a compact subset of $\mathcal{O}$, show that the mapping $\mathbf{F} : C \to \mathbb{R}^n$ is stable.

5.3 The Implicit Function Theorem

The implicit function theorem is about the possibility of solving m variables from a system of m equations with $n + m$ variables. Let us study some special cases.

Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by $f(x, y) = x^2 + y^2 - 1$. For a point (x_0, y_0) that satisfies $f(x_0, y_0) = 0$, we want to ask whether there is a neighbourhood I of x_0, a neighbourhood J of y_0, and a function $g : I \to \mathbb{R}$ such that for $(x, y) \in I \times J$, $f(x, y) = 0$ if and only if $y = g(x)$.

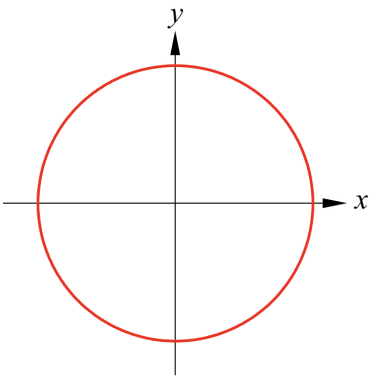


Figure 5.4: The points in the (x, y) plane satisfying $x^2 + y^2 - 1 = 0$.

If (x_0, y_0) is a point with $y_0 > 0$ and $f(x_0, y_0) = 0$, then we can take the neighbourhoods $I = (-1, 1)$ and $J = (0, \infty)$ of x_0 and y_0 respectively, and define the function $g : I \to \mathbb{R}$ by

$$g(x) = \sqrt{1 - x^2}.$$

We then find that for $(x, y) \in I \times J$, $f(x, y) = 0$ if and only if $y = \sqrt{1 - x^2} = g(x)$.

If (x_0, y_0) is a point with $y_0 < 0$ and $f(x_0, y_0) = 0$, then we can take the neighbourhoods $I = (-1, 1)$ and $J = (-\infty, 0)$ of x_0 and y_0 respectively, and define the function $g : I \to \mathbb{R}$ by

$$g(x) = -\sqrt{1 - x^2}.$$

We then find that for $(x, y) \in I \times J$, $f(x, y) = 0$ if and only if $y = -\sqrt{1 - x^2} = g(x)$.

However, if $(x_0, y_0) = (1, 0)$, any neighbourhood J of y_0 must contain an interval of the form $(-r, r)$. If I is a neighbourhood of 1, (x, y) is a point in

$I \times (-r, r)$ such that $f(x, y) = 0$, then $(x, -y)$ is another point in $I \times (-r, r)$ satisfying $f(x, -y) = 0$. This shows that there does not exist any function $g : I \to \mathbb{R}$ such that when $(x, y) \in I \times J$, $f(x, y) = 0$ if and only if $y = g(x)$. We say that we cannot solve y as a function of x in a neighbourhood of the point $(1, 0)$.

Similarly, we cannot solve y as a function of x in a neighbourhood of the point $(-1, 0)$.

However, in a neighbourhood of the points $(1, 0)$ and $(-1, 0)$, we can solve x as a function of y.

For a function $f : \mathcal{O} \to \mathbb{R}$ defined on an open subset $\mathcal{O}$ of $\mathbb{R}^2$, the implicit function theorem takes the following form.

Theorem 5.9 Dini's Theorem

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^2$ that contains the point (x_0, y_0), and let $f : \mathcal{O} \to \mathbb{R}$ be a continuously differentiable function defined on $\mathcal{O}$ such that $f(x_0, y_0) = 0$. If $\dfrac{\partial f}{\partial y}(x_0, y_0) \neq 0$, then there is a neighbourhood I of x_0, a neighbourhood J of y_0, and a continuously differentiable function $g : I \to J$ such that for any $(x, y) \in I \times J$, $f(x, y) = 0$ if and only if $y = g(x)$. Moreover, for any $x \in I$,

$$\frac{\partial f}{\partial x}(x, g(x)) + \frac{\partial f}{\partial y}(x, g(x))g'(x) = 0.$$

Dini's theorem says that to be able to solve y as a function of x, a sufficient condition is that the function f has continuous partial derivatives, and f_y does not vanish. By interchanging the roles of x and y, we see that if f_x does not vanish, we can solve x as a function of y.

For the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = x^2 + y^2 - 1$, the points on the set $x^2 + y^2 = 1$ which $f_y(x, y) = 2y$ vanishes are the points $(1, 0)$ and $(-1, 0)$. In fact, we have seen that we cannot solve y as functions of x in neighbourhoods of these two points.

Proof of Dini's Theorem

Without loss of generality, assume that $f_y(x_0, y_0) > 0$. Let $\mathbf{u}_0 = (x_0, y_0)$. Since $f_y : \mathcal{O} \to \mathbb{R}$ is continuous, there is an $r_1 > 0$ such that the closed rectangle $R = [x_0 - r_1, x_0 + r_1] \times [y_0 - r_1, y_0 + r_1]$ lies in $\mathcal{O}$, and for all $(x, y) \in R$, $f_y(x, y) > f_y(x_0, y_0)/2 > 0$. For any $x \in [x_0 - r_1, x_0 + r_1]$, the function $h_x : [y_0 - r_1, y_0 + r_1] \to \mathbb{R}$ has derivative $h'_x(y) = f_y(x, y)$ that is positive. Hence, $h_x(y) = g(x, y)$ is strictly increasing in y. This implies that

$$f(x, y_0 - r_1) < f(x, y_0) < f(x, y_0 + r_1).$$

When $x = x_0$, we find that

$$f(x_0, y_0 - r_1) < 0 < f(x_0, y_0 + r_1).$$

Since f is continuously differentiable, it is continuous. Hence, there is an $r_2 > 0$ such that $r_2 \leq r_1$, and for all $x \in [x_0 - r_2, x_0 + r_2]$,

$$f(x, y_0 - r_1) < 0 \quad \text{and} \quad f(x, y_0 + r_1) > 0.$$

Let $I = (x_0 - r_2, x_0 + r_2)$. For $x \in I$, since $h_x : [y_0 - r_1, y_0 + r_1] \to \mathbb{R}$ is continuous, and

$$h_x(y_0 - r_1) < 0 < h_x(y_0 + r_1),$$

intermediate value theorem implies that there is a $y \in (y_0 - r_1, y_0 + r_1)$ such that $h_x(y) = 0$. Since h_x is strictly increasing, this y is unique, and we denote it by $g(x)$. This defines the function $g : I \to \mathbb{R}$. Let $J = (y_0 - r_1, y_0 + r_1)$. By our argument, for each $x \in I$, $y = g(x)$ is a unique $y \in J$ such that $f(x, y) = 0$. Thus, for any $(x, y) \in I \times J$, $f(x, y) = 0$ if and only if $y = g(x)$.

It remains to prove that $g : I \to \mathbb{R}$ is continuosly differentiable. By our convention above, there is a positive constant c such that

$$\frac{\partial f}{\partial y}(x, y) \geq c \qquad \text{for all } (x, y) \in I \times J.$$

Fixed $x \in I$. There exists an $r > 0$ such that $(x - r, x + r) \subset I$. For h satisfying $0 < |h| < r$, $x + h$ is in I. By mean value theorem, there is a $c_h \in (0, 1)$ such that

$$f(x+h, g(x+h)) - f(x, g(x)) = h\frac{\partial f}{\partial x}(\mathbf{u}_h) + (g(x+h) - g(x))\frac{\partial f}{\partial y}(\mathbf{u}_h),$$

where

$$\mathbf{u}_h = (x, g(x)) + c_h(h, g(x+h) - g(x)). \tag{5.4}$$

Since

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

we find that

$$\frac{g(x+h) - g(x)}{h} = -\frac{f_x(\mathbf{u}_h)}{f_y(\mathbf{u}_h)}. \tag{5.5}$$

Since f_x is continuous on the compact set R, it is bounded. Namely, there exists a constant M such that

$$|f_x(x, y)| \leq M \qquad \text{for all } (x, y) \in R.$$

Eq. (5.5) then implies that

$$|g(x+h) - g(x)| \leq \frac{M}{c}|h|.$$

Taking $h \to 0$ proves that g is continuous at x. From (5.4), we find that

$$\lim_{h\to 0} \mathbf{u}_h = (x, g(x)).$$

Since f_x and f_y are continuous at $(x, g(x))$, eq. (5.5) gives

$$\lim_{h\to 0} \frac{g(x+h) - g(x)}{h} = -\lim_{h\to 0} \frac{f_x(\mathbf{u}_h)}{f_y(\mathbf{u}_h)} = -\frac{f_x(x, g(x))}{f_y(x, g(x))}.$$

This proves that g is differentiable at x and

$$\frac{\partial f}{\partial x}(x, g(x)) + \frac{\partial f}{\partial y}(x, g(x))g'(x) = 0.$$

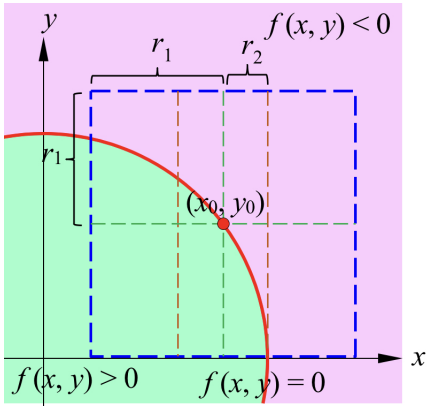


Figure 5.5: Proof of Dini's Theorem.

Example 5.8

Consider the equation

$$xy^3 + \sin(x + y) + 4x^2y = 3.$$

Show that in a neighbourhood of $(-1, 1)$, this equation defines y as a function of x. If this function is denoted as $y = g(x)$, find $g'(-1)$.

Solution

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function defined as

$$f(x, y) = xy^3 + \sin(x + y) + 4x^2y - 3.$$

Since sine function and polynomial functions are infinitely differentiable, f is infinitely differentiable.

$$\frac{\partial f}{\partial y}(x, y) = 3xy^2 + \cos(x + y) + 4x^2, \qquad \frac{\partial f}{\partial y}(-1, 1) = 2 \neq 0.$$

By Dini's theorem, there is a neighbourhood of $(-1, 1)$ such that y can be solved as a function of x. Now,

$$\frac{\partial f}{\partial x}(x, y) = y^3 + \cos(x + y) + 8xy, \qquad \frac{\partial f}{\partial x}(-1, 1) = -6.$$

Hence, $g'(-1) = -\dfrac{-6}{2} = 3.$

Now we turn to the general case. First we consider polynomial mappings of degree at most one. Let $A = [a_{ij}]$ be an $m \times n$ matrix, and let $B = [b_{ij}]$ be an $m \times m$ matrix. Given $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$, $\mathbf{c} \in \mathbb{R}^m$, the system of equations

$$A\mathbf{x} + B\mathbf{y} = \mathbf{c}$$

is the following m equations in $m + n$ variables $x_1, \ldots, x_n, y_1, \ldots, y_m$.

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n + b_{11}y_1 + b_{12}y_2 + \cdots + b_{1m}y_m &= c_1,\\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n + b_{21}y_1 + b_{22}y_2 + \cdots + b_{2m}y_m &= c_2,\\
&\vdots\\
a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n + b_{m1}y_1 + b_{m2}y_2 + \cdots + b_{mm}y_m &= c_m.
\end{aligned}$$

Let us look at an example.

Example 5.9

Consider the linear system

$$\begin{aligned}
2x_1 + 3x_2 - 5x_3 + 2y_1 - y_2 &= 1\\
3x_1 - x_2 + 2x_3 - 3y_1 + y_2 &= 0
\end{aligned}$$

Show that $\mathbf{y} = (y_1, y_2)$ can be solved as a function of $\mathbf{x} = (x_1, x_2, x_3)$. Write down the function $\mathbf{G} : \mathbb{R}^3 \to \mathbb{R}^2$ such that the solution is given by $\mathbf{y} = \mathbf{G}(\mathbf{x})$, and find $\mathbf{DG}(\mathbf{x})$.

Solution

Let

$$A = \begin{bmatrix} 2 & 3 & -5 \\ 3 & -1 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & -1 \\ -3 & 1 \end{bmatrix}.$$

Then the system can be written as

$$A\mathbf{x} + B\mathbf{y} = \mathbf{c}, \qquad \text{where } \mathbf{c} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

This implies that

$$B\mathbf{y} = \mathbf{c} - A\mathbf{x}. \tag{5.6}$$

For every $\mathbf{x} \in \mathbb{R}^3$, $\mathbf{c} - A\mathbf{x}$ is a vector in $\mathbb{R}^2$. Since $\det B = -1 \neq 0$, B is invertible. Therefore, there is a unique $\mathbf{y}$ satisfying (5.6). It is given by

$$\begin{aligned}\mathbf{G}(\mathbf{x}) = \mathbf{y} &= B^{-1}(\mathbf{c} - A\mathbf{x}) \\ &= -\begin{bmatrix}1 & 1\\ 3 & 2\end{bmatrix}\begin{bmatrix}1\\ 0\end{bmatrix} + \begin{bmatrix}1 & 1\\ 3 & 2\end{bmatrix}\begin{bmatrix}2 & 3 & -5\\ 3 & -1 & 2\end{bmatrix}\mathbf{x} \\ &= -\begin{bmatrix}1\\ 3\end{bmatrix} + \begin{bmatrix}5 & 2 & -3\\ 12 & 7 & -11\end{bmatrix}\mathbf{x} \\ &= \begin{bmatrix}5x_1 + 2x_2 - 3x_3 - 1\\ 12x_1 + 7x_2 - 11x_3 - 3\end{bmatrix}.\end{aligned}$$

It follows that $\mathbf{DG} = \begin{bmatrix}5 & 2 & -3\\ 12 & 7 & -11\end{bmatrix}$.

The following theorem gives a general scenario.

Theorem 5.10

Let $A = [a_{ij}]$ be an $m \times n$ matrix, and let $B = [b_{ij}]$ be an $m \times m$ matrix. Define the function $\mathbf{F} : \mathbb{R}^{m+n} \to \mathbb{R}^m$ by

$$\mathbf{F}(\mathbf{x}, \mathbf{y}) = A\mathbf{x} + B\mathbf{y} - \mathbf{c},$$

where $\mathbf{c}$ is a constant vector in $\mathbb{R}^m$. The equation $\mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ defines the variable $\mathbf{y} = (y_1, \ldots, y_m)$ as a function of $\mathbf{x} = (x_1, \ldots, x_n)$ if and only if the matrix B is invertible. If we denote this function as $\mathbf{G} : \mathbb{R}^n \to \mathbb{R}^m$, then

$$\mathbf{G}(\mathbf{x}) = B^{-1}(\mathbf{c} - A\mathbf{x}),$$

and

$$\mathbf{DG}(\mathbf{x}) = -B^{-1}A.$$

Proof

The equation $\mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ defines the variables $\mathbf{y}$ as a function of $\mathbf{x}$ if and only for for each $\mathbf{x} \in \mathbb{R}^n$, there is a unique $\mathbf{y} \in \mathbb{R}^m$ satisfying

$$B\mathbf{y} = \mathbf{c} - A\mathbf{x}.$$

This is a linear system for the variable $\mathbf{y}$. By the theory of linear algebra, a unique solution $\mathbf{y}$ exists if and only if B is invertible. In this case, the solution is given by

$$\mathbf{y} = B^{-1}(\mathbf{c} - A\mathbf{x}).$$

The rest of the assertion follows.

Write a point in $\mathbb{R}^{m+n}$ as $(\mathbf{x}, \mathbf{y})$, where $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$. If $\mathbf{F} : \mathbb{R}^{m+n} \to \mathbb{R}^m$ is a function that is differentiable at the point $(\mathbf{x}, \mathbf{y})$, the $m \times (m+n)$ derivative matrix $\mathbf{DF}(\mathbf{x}, \mathbf{y})$ can be written as

$$\mathbf{DF}(\mathbf{x}, \mathbf{y}) = \left[\mathbf{D_x F}(\mathbf{x}, \mathbf{y}) \;\middle|\; \mathbf{D_y F}(\mathbf{x}, \mathbf{y}) \right],$$

where

$$\mathbf{D_x F}(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} \dfrac{\partial F_1}{\partial x_1}(\mathbf{x}, \mathbf{y}) & \dfrac{\partial F_1}{\partial x_2}(\mathbf{x}, \mathbf{y}) & \cdots & \dfrac{\partial F_1}{\partial x_n}(\mathbf{x}, \mathbf{y}) \\ \dfrac{\partial F_2}{\partial x_1}(\mathbf{x}, \mathbf{y}) & \dfrac{\partial F_2}{\partial x_2}(\mathbf{x}, \mathbf{y}) & \cdots & \dfrac{\partial F_2}{\partial x_n}(\mathbf{x}, \mathbf{y}) \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial F_m}{\partial x_1}(\mathbf{x}, \mathbf{y}) & \dfrac{\partial F_m}{\partial x_2}(\mathbf{x}, \mathbf{y}) & \cdots & \dfrac{\partial F_m}{\partial x_n}(\mathbf{x}, \mathbf{y}) \end{bmatrix},$$

$$\mathbf{D_y F}(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} \dfrac{\partial F_1}{\partial y_1}(\mathbf{x}, \mathbf{y}) & \dfrac{\partial F_1}{\partial y_2}(\mathbf{x}, \mathbf{y}) & \cdots & \dfrac{\partial F_1}{\partial y_m}(\mathbf{x}, \mathbf{y}) \\ \dfrac{\partial F_2}{\partial y_1}(\mathbf{x}, \mathbf{y}) & \dfrac{\partial F_2}{\partial y_2}(\mathbf{x}, \mathbf{y}) & \cdots & \dfrac{\partial F_2}{\partial y_m}(\mathbf{x}, \mathbf{y}) \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial F_m}{\partial y_1}(\mathbf{x}, \mathbf{y}) & \dfrac{\partial F_m}{\partial y_2}(\mathbf{x}, \mathbf{y}) & \cdots & \dfrac{\partial F_m}{\partial y_m}(\mathbf{x}, \mathbf{y}) \end{bmatrix}.$$

Notice that $\mathbf{D_y F}(\mathbf{x}, \mathbf{y})$ is a square matrix.

When $A = [a_{ij}]$ is an $m \times n$ matrix, $B = [b_{ij}]$ is an $m \times m$ matrix, $\mathbf{c}$ is a vector in $\mathbb{R}^m$, and $\mathbf{F} : \mathbb{R}^{m+n} \to \mathbb{R}^m$ is the function defined as

$$\mathbf{F}(\mathbf{x}, \mathbf{y}) = A\mathbf{x} + B\mathbf{y} - \mathbf{c},$$

it is easy to compute that

$$\mathbf{D_x F}(\mathbf{x}, \mathbf{y}) = A, \qquad \mathbf{D_y F}(\mathbf{x}, \mathbf{y}) = B.$$

Theorem 5.10 says that we can solve $\mathbf{y}$ as a function of $\mathbf{x}$ from the system of m equations

$$\mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$$

if and only if

$$B = \mathbf{D_y F}(\mathbf{x}, \mathbf{y})$$

is invertible. In this case, if $\mathbf{G} : \mathbb{R}^n \to \mathbb{R}^m$ is the function so that $\mathbf{y} = \mathbf{G}(\mathbf{x})$ is the solution, then

$$\mathbf{DG}(\mathbf{x}) = -B^{-1}A = -\mathbf{D_y F}(\mathbf{x}, \mathbf{y})^{-1}\mathbf{D_x F}(\mathbf{x}, \mathbf{y}).$$

In fact, this latter follows from $\mathbf{F}(\mathbf{x}, \mathbf{G}(\mathbf{x})) = \mathbf{0}$ and the chain rule.

The special case of degree one polynomial mappings gives us sufficient insight into the general implicit function theorem. However, for nonlinear mappings, the conclusions can only be made *locally*.

Theorem 5.11 Implicit Function Theorem

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^{m+n}$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ be a continuously differentiable function defined on $\mathcal{O}$. Assume that $\mathbf{x}_0$ is a point in $\mathbb{R}^n$ and $\mathbf{y}_0$ is a point in $\mathbb{R}^m$ such that the point $(\mathbf{x}_0, \mathbf{y}_0)$ is in $\mathcal{O}$ and $\mathbf{F}(\mathbf{x}_0, \mathbf{y}_0) = \mathbf{0}$. If $\det \mathbf{D_y F}(\mathbf{x}_0, \mathbf{y}_0) \neq 0$, then we have the followings.

(i) There is a neighbourhood U of $\mathbf{x}_0$, a neighbourhood V of $\mathbf{y}_0$, and a continuously differentiable function $\mathbf{G} : U \to \mathbb{R}^m$ such that for any $(\mathbf{x}, \mathbf{y}) \in U \times V$, $\mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ if and only if $\mathbf{y} = \mathbf{G}(\mathbf{x})$.

(ii) For any $\mathbf{x} \in U$,

$$\mathbf{D_x F}(\mathbf{x}, \mathbf{G}(\mathbf{x})) + \mathbf{D_y F}(\mathbf{x}, \mathbf{G}(\mathbf{x}))\mathbf{DG}(\mathbf{x}) = \mathbf{0}.$$

Here we will give a proof of the implicit function theorem using the inverse function theorem. The idea of the proof is to construct a mapping which one can apply the inverse function theorem. Let us look at an example first.

Example 5.10

Let $\mathbf{F} : \mathbb{R}^5 \to \mathbb{R}^2$ be the function defined as

$$\mathbf{F}(x_1, x_3, x_3, y_1, y_2) = (x_1 y_2^2, x_2 x_3 y_1^2 + x_1 y_2).$$

Define the mapping $\mathbf{H} : \mathbb{R}^5 \to \mathbb{R}^5$ as

$$\mathbf{H}(\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathbf{F}(\mathbf{x}, \mathbf{y})) = (x_1, x_2, x_3, x_1 y_2^2, x_2 x_3 y_1^2 + x_1 y_2).$$

Then we find that

$$\mathbf{DH}(\mathbf{x}, \mathbf{y}) = \left[\begin{array}{ccc|cc} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \hline y_2^2 & 0 & 0 & 0 & 2x_1 y_2 \\ y_2 & x_3 y_1^2 & x_2 y_1^2 & 2x_2 x_3 y_1 & x_1 \end{array}\right].$$

Notice that

$$\mathbf{DH}(\mathbf{x}, \mathbf{y}) = \left[\begin{array}{c|c} I_3 & \mathbf{0} \\ \hline \mathbf{D_x F}(\mathbf{x}, \mathbf{y}) & \mathbf{D_y F}(\mathbf{x}, \mathbf{y}) \end{array}\right].$$

Proof of the Implicit Function Theorem

Let $\mathbf{H} : \mathcal{O} \to \mathbb{R}^{m+n}$ be the mapping defined as

$$\mathbf{H}(\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathbf{F}(\mathbf{x}, \mathbf{y})).$$

Notice that $\mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ if and only if $\mathbf{H}(\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathbf{0})$. Since the first n components of $\mathbf{H}$ are infinitely differentiable functions, the mapping $\mathbf{H} : \mathcal{O} \to \mathbb{R}^{m+n}$ is continuously differentiable.

Now,

$$\mathbf{DH}(\mathbf{x},\mathbf{y}) = \left[\begin{array}{c|c} I_n & \mathbf{0} \\ \hline \mathbf{D_x F}(\mathbf{x},\mathbf{y}) & \mathbf{D_y F}(\mathbf{x},\mathbf{y}) \end{array}\right].$$

Therefore,

$$\det \mathbf{DH}(\mathbf{x}_0,\mathbf{y}_0) = \det \mathbf{D_y F}(\mathbf{x}_0,\mathbf{y}_0) \neq 0.$$

By the inverse function theorem, there is a neighbourhood W of $(\mathbf{x}_0,\mathbf{y}_0)$ and a neighbourhood Z of $\mathbf{H}(\mathbf{x}_0,\mathbf{y}_0) = (\mathbf{x}_0,\mathbf{0})$ such that $\mathbf{H} : W \to Z$ is a bijection and $\mathbf{H}^{-1} : Z \to W$ is continuously differentiable. For $\mathbf{u} \in \mathbb{R}^n$, $\mathbf{v} \in \mathbb{R}^m$ so that $(\mathbf{u},\mathbf{v}) \in Z$, let

$$\mathbf{H}^{-1}(\mathbf{u},\mathbf{v}) = (\mathbf{\Phi}(\mathbf{u},\mathbf{v}), \mathbf{\Psi}(\mathbf{u},\mathbf{v})),$$

where $\mathbf{\Phi}$ is a map from Z to $\mathbb{R}^n$ and $\mathbf{\Psi}$ is a map from Z to $\mathbb{R}^m$. Since $\mathbf{H}^{-1}$ is continuously differentiable, $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are continuously differentiable.

Given $r > 0$, let D_r be the open cube $D_r = \prod_{i=1}^{m+n} (-r,r)$. Since W and Z are open sets that contain $(\mathbf{x}_0,\mathbf{y}_0)$ and $(\mathbf{x}_0,\mathbf{0})$ respectively, there exists $r > 0$ such that

$$(\mathbf{x}_0,\mathbf{y}_0) + D_r \subset W, \qquad (\mathbf{x}_0,\mathbf{0}) + D_r \subset Z.$$

If $A_r = \prod_{i=1}^{n}(-r,r)$, $B_r = \prod_{i=1}^{m}(-r,r)$, $U = \mathbf{x}_0 + A_r$, $V = \mathbf{y}_0 + B_r$, then

$$(\mathbf{x}_0,\mathbf{y}_0) + D_r = U \times V, \qquad (\mathbf{x}_0,\mathbf{0}) + D_r = U \times B_r.$$

Hence, $U \times V \subset W$ and $U \times B_r \subset Z$. Define $\mathbf{G} : U \to \mathbb{R}^m$ by

$$\mathbf{G}(\mathbf{x}) = \mathbf{\Psi}(\mathbf{x},\mathbf{0}).$$

Since $\mathbf{\Psi}$ is continuously differentiable, $\mathbf{G}$ is continuously differentiable. If $\mathbf{x} \in U$, $\mathbf{y} \in V$, then $(\mathbf{x},\mathbf{y}) \in W$. For such $(\mathbf{x},\mathbf{y})$, $\mathbf{F}(\mathbf{x},\mathbf{y}) = \mathbf{0}$ implies $\mathbf{H}(\mathbf{x},\mathbf{y}) = (\mathbf{x},\mathbf{0})$. Since $\mathbf{H} : W \to Z$ is a bijection, $(\mathbf{x},\mathbf{0}) \in Z$ and $\mathbf{H}^{-1}(\mathbf{x},\mathbf{0}) = (\mathbf{x},\mathbf{y})$. Comparing the last m components give

$$\mathbf{y} = \mathbf{\Psi}(\mathbf{x},\mathbf{0}) = \mathbf{G}(\mathbf{x}).$$

Conversely, since $\mathbf{H}(\mathbf{H}^{-1}(\mathbf{u},\mathbf{v})) = (\mathbf{u},\mathbf{v})$ for all $(\mathbf{u},\mathbf{v}) \in Z$, we find that

$$(\mathbf{\Phi}(\mathbf{u},\mathbf{v}), \mathbf{F}(\mathbf{\Phi}(\mathbf{u},\mathbf{v}), \mathbf{\Psi}(\mathbf{u},\mathbf{v}))) = (\mathbf{u},\mathbf{v})$$

for all $(\mathbf{u},\mathbf{v}) \in Z$. For all $\mathbf{u} \in U$, $(\mathbf{u},\mathbf{0})$ is in Z. Therefore,

$$\mathbf{\Phi}(\mathbf{u},\mathbf{0}) = \mathbf{u}, \quad \mathbf{F}(\mathbf{\Phi}(\mathbf{u},\mathbf{0}), \mathbf{\Psi}(\mathbf{u},\mathbf{0})) = \mathbf{0}.$$

This implies that if $\mathbf{x} \in U$, then $\mathbf{F}(\mathbf{u}, \mathbf{G}(\mathbf{u})) = \mathbf{0}$. In other words, if $(\mathbf{x},\mathbf{y})$ is in $U \times V$ and $\mathbf{y} = \mathbf{G}(\mathbf{x})$, we must have $\mathbf{F}(\mathbf{x},\mathbf{y}) = \mathbf{0}$. Since we have shown that $\mathbf{G} : U \to \mathbb{R}^m$ is continuously differentiable, the formula

$$\mathbf{D_x F}(\mathbf{x}, \mathbf{G}(\mathbf{x})) + \mathbf{D_y F}(\mathbf{x}, \mathbf{G}(\mathbf{x}))\mathbf{DG}(\mathbf{x}) = \mathbf{0}$$

follows from $\mathbf{F}(\mathbf{x}, \mathbf{G}(\mathbf{x})) = \mathbf{0}$ and the chain rule.

Example 5.11

Consider the system of equations

$$\begin{aligned} 2x^2y + 3xy^2u + xyv + uv &= 7 \\ 4xu - 5yv + u^2y + v^2x &= 1 \end{aligned} \tag{5.7}$$

Notice that when $(x,y) = (1,1)$, $(u,v) = (1,1)$ is a solution of this system. Show that there are neighbourhoods U and V of $(1,1)$, and a continuously differentiable function $\mathbf{G} : U \to \mathbb{R}^2$ such that if $(x,y,u,v) \in U \times V$, then (x,y,u,v) is a solution of the system of equations above if and only if $u = G_1(x,y)$ and $v = G_2(x,y)$. Also, find the values of $\dfrac{\partial G_1}{\partial x}(1,1)$, $\dfrac{\partial G_1}{\partial y}(1,1)$, $\dfrac{\partial G_2}{\partial x}(1,1)$ and $\dfrac{\partial G_2}{\partial y}(1,1)$.

Solution

Define the function $\mathbf{F} : \mathbb{R}^4 \to \mathbb{R}^2$ by

$$\mathbf{F}(x,y,u,v) = (2x^2y + 3xy^2u + xyv + uv - 7, 4xu - 5yv + u^2y + v^2x - 1).$$

This is a polynomial mapping. Hence, it is continuously differentiable. It is easy to check that $\mathbf{F}(1,1,1,1) = \mathbf{0}$. Now,

$$\mathbf{D}_{(u,v)}\mathbf{F}(x,y,u,v) = \begin{bmatrix} 3xy^2 + v & xy + u \\ 4x + 2uy & -5y + 2vx \end{bmatrix}.$$

Thus,

$$\det \mathbf{D}_{(u,v)}\mathbf{F}(1,1,1,1) = \begin{bmatrix} 4 & 2 \\ 6 & -3 \end{bmatrix} = -24 \neq 0.$$

By implicit function theorem, there are neighbourhoods U and V of $(1,1)$, and a continuously differentiable function $\mathbf{G} : U \to \mathbb{R}^2$ such that, if $(x,y,u,v) \in U \times V$, then (x,y,u,v) is a solution of the system of equations (5.7) if and only if $u = G_1(x,y)$ and $v = G_2(x,y)$.
Finally,

$$\mathbf{D}_{(x,y)}\mathbf{F}(x,y,u,v) = \begin{bmatrix} 4xy + 3y^2u + yv & 2x^2 + 6xyu + xv \\ 4u + v^2 & -5v + u^2 \end{bmatrix},$$

$$\mathbf{D}_{(x,y)}\mathbf{F}(1,1,1,1) = \begin{bmatrix} 8 & 9 \\ 5 & -4 \end{bmatrix}.$$

Chain rule gives

$$\begin{aligned} \mathbf{DG}(1,1) &= -\mathbf{D}_{(u,v)}\mathbf{F}(1,1,1,1)^{-1}\mathbf{D}_{(x,y)}\mathbf{F}(1,1,1,1) \\ &= \frac{1}{24}\begin{bmatrix} -3 & -2 \\ -6 & 4 \end{bmatrix}\begin{bmatrix} 8 & 9 \\ 5 & -4 \end{bmatrix} \\ &= \frac{1}{24}\begin{bmatrix} -34 & -19 \\ -28 & -70 \end{bmatrix}. \end{aligned}$$

Therefore,

$$\frac{\partial G_1}{\partial x}(1,1) = -\frac{17}{12}, \qquad \frac{\partial G_1}{\partial y}(1,1) = -\frac{19}{24},$$

$$\frac{\partial G_2}{\partial x}(1,1) = -\frac{7}{6}, \qquad \frac{\partial G_2}{\partial y}(1,1) = -\frac{35}{12}.$$

Remark 5.3 The Rank of a Matrix

In the formulation of the implicit function theorem, the assumption that $\det \mathbf{D_y F}(\mathbf{x}_0, \mathbf{y}_0) \neq 0$ can be replaced by the assumption that there are m variables $u_1, \ldots, u_m$ among the $n+m$ variables $x_1, \ldots, x_n, y_1, \ldots, y_m$ such that $\det \mathbf{D}_{(u_1,\ldots,u_m)}\mathbf{F}(\mathbf{x}_0, \mathbf{y}_0) \neq 0$.

Recall that the rank r of an $m \times k$ matrix A is the dimension of its row space or the dimension of its column space. Thus, the rank r of a $m \times k$ matrix A is the maximum number of column vectors of A which are linearly independent, or the maximum number of row vectors of A that are linearly independent. Hence, the maximum possible value of r is $\max\{m, k\}$. If $r = \max\{m, k\}$, we say that the matrix A has maximal rank. For a $m \times k$ matrix where $m \leq k$, it has maximal rank if $r = m$. In this case, there is a $m \times m$ submatrix of A consists of m linearly independent vectors in $\mathbb{R}^m$. The determinant of this submatrix is nonzero.

Thus, the condition $\det \mathbf{D_y F}(\mathbf{x}_0, \mathbf{y}_0) \neq 0$ in the formulation of the implicit function theorem can be replaced by the condition that the $m \times (m + n)$ matrix $\mathbf{DF}(\mathbf{x}_0, \mathbf{y}_0)$ has maximal rank.

Example 5.12

Consider the system

$$\begin{aligned} 2x^2y + 3xy^2u + xyv + uv &= 7 \\ 4xu - 5yv + u^2y + v^2x &= 1 \end{aligned} \tag{5.8}$$

defined in Example 5.11. Show that there are neighbourhoods U and V of $(1, 1)$, and a continuously differentiable function $\mathbf{H} : V \to \mathbb{R}^2$ such that if $(x, y, u, v) \in U \times V$, then (x, y, u, v) is a solution of the system of equations if and only if $x = H_1(u, v)$ and $y = H_2(u, v)$. Find $\mathbf{DH}(1, 1)$.

Solution

Define the function $\mathbf{F} : \mathbb{R}^4 \to \mathbb{R}^2$ as in the solution of Example 5.11. Since

$$\det \mathbf{D}_{(x,y)}\mathbf{F}(1,1,1,1) = \begin{bmatrix} 8 & 9 \\ 5 & -4 \end{bmatrix} = -77 \neq 0,$$

the implicit function theorem implies there are neighbourhoods U and V of $(1,1)$, and a continuously differentiable function $\mathbf{H} : V \to \mathbb{R}^2$ such that if $(x,y,u,v) \in U \times V$, then (x,y,u,v) is a solution of the system of equations (5.8) if and only if $x = H_1(u,v)$ and $y = H_2(u,v)$. Moreover,

$$\begin{aligned} \mathbf{DH}(1,1) &= -\mathbf{D}_{(x,y)}\mathbf{F}(1,1,1,1)^{-1}\mathbf{D}_{(u,v)}\mathbf{F}(1,1,1,1) \\ &= \frac{1}{77}\begin{bmatrix} -4 & -9 \\ -5 & 8 \end{bmatrix}\begin{bmatrix} 4 & 2 \\ 6 & -3 \end{bmatrix} \\ &= \frac{1}{77}\begin{bmatrix} -70 & 19 \\ 28 & -34 \end{bmatrix}. \end{aligned}$$

Remark 5.4

The function $\mathbf{G} : U \to \mathbb{R}^2$ in Example 5.11 and the function $\mathbf{H} : V \to \mathbb{R}^2$ in Example 5.12 are in fact inverses of each other.

Notice that $\mathbf{DG}(1,1)$ is invertible. By the inverse function theorem, there is a neighbourhood U' of $(1,1)$ such that $V' = \mathbf{G}(U)$ is open, and $\mathbf{G} : U' \to V'$ is a bijection with continuously differentiable inverse. By shrinking down the sets U and V, we can assume that $U = U'$, and $V = V'$. If $(x,y) \in U$ and $(u,v) \in V$, $\mathbf{F}(x,y,u,v) = 0$ if and only if $(u,v) = \mathbf{G}(x,y)$, if and only if $(x,y) = \mathbf{H}(u,v)$. This implies that $\mathbf{G} : U \to V$ and $\mathbf{H} : V \to U$ are inverses of each other.

At the end of this section, let us consider a geometric application of the implicit function theorem. First let us revisit the example where $f(x,y) = x^2 + y^2 - 1$. At each point (x_0, y_0) such that $f(x_0, y_0) = 0$,

$$x_0^2 + y_0^2 = 1.$$

Hence, $\nabla f(x_0, y_0) = (2x_0, 2y_0) \neq \mathbf{0}$. Notice that the vector $\nabla f(x_0, y_0) =$

$(2x_0, 2y_0)$ is normal to the circle $x^2 + y^2 = 1$ at the point (x_0, y_0).

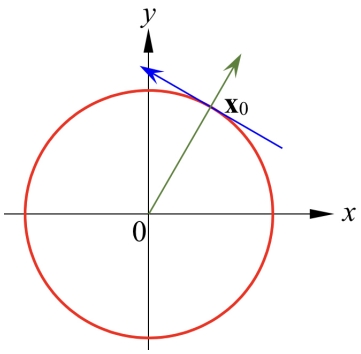


Figure 5.6: The tangent vector and normal vector at a point on the circle $x^2 + y^2 - 1 = 0$.

If $y_0 > 0$, let $U = (-1, 1) \times (0, \infty)$. Restricted to U, the points where $f(x, y) = 0$ is the graph of the function $g : (-1, 1) \to \mathbb{R}$, $g(x) = \sqrt{1 - x^2}$.

If $y_0 < 0$, let $U = (-1, 1) \times (-\infty, 0)$. Restricted to U, the points where $f(x, y) = 0$ is the graph of the function $g : (-1, 1) \to \mathbb{R}$, $g(x) = -\sqrt{1 - x^2}$.

If $y_0 = 0$, then $x_0 = 1$ or -1. In fact, we can consider more generally the cases where $x_0 > 0$ and $x_0 < 0$.

If $x_0 > 0$, let $U = (0, \infty) \times (-1, 1)$. Restricted to U, the points where $f(x, y) = 0$ is the graph of the function $g : (-1, 1) \to \mathbb{R}$, $g(y) = \sqrt{1 - y^2}$.

If $x_0 < 0$, let $U = (-\infty, 0) \times (-1, 1)$. Restricted to U, the points where $f(x, y) = 0$ is the graph of the function $g : (-1, 1) \to \mathbb{R}$, $g(y) = -\sqrt{1 - y^2}$.

Definition 5.4 Surfaces

Let S be a subset of $\mathbb{R}^k$ for some positive integer k. We say that S is a n-dimensional surface if for each $\mathbf{x}_0$ on S, there is an open subset $\mathcal{D}$ of $\mathbb{R}^n$, an open neighbourhood $\mathcal{U}$ of $\mathbf{x}_0$ in $\mathbb{R}^k$, and a one-to-one differentiable mapping $\mathbf{G} : \mathcal{D} \to \mathbb{R}^k$ such that $\mathbf{G}(\mathcal{D}) \subset S$, $\mathbf{G}(\mathcal{D}) \cap \mathcal{U} = S \cap \mathcal{U}$, and $\mathbf{DG}(\mathbf{u})$ has rank n at each $\mathbf{u} \in \mathcal{D}$.

Example 5.13

We claim that the n-sphere

$$S^n = \{(x_1, \ldots, x_n, x_{n+1}) \,|\, x_1^2 + \cdots + x_n^2 + x_{n+1}^2 = 1\}$$

is an n-dimensional surface. Let $(a_1, \ldots, a_n, a_{n+1})$ be a point on S^n. Then at least one of the components $a_1, \ldots, a_n, a_{n+1}$ is nonzero. Without loss of generality, assume that $a_{n+1} > 0$. Let

$$\mathcal{D} = \left\{(x_1, \ldots, x_n) \,|\, x_1^2 + \cdots + x_n^2 < 1\right\},$$

$$\mathcal{U} = \{(x_1, \ldots, x_n, x_{n+1}) \,|\, x_{n+1} > 0\},$$

and define the mapping $\mathbf{G} : \mathcal{D} \to \mathcal{U}$ by

$$\mathbf{G}(x_1, \ldots, x_n) = \left(x_1, \ldots, x_n, \sqrt{1 - x_1^2 - \cdots - x_n^2}\right).$$

Then $\mathbf{G}$ is a differentiable mapping, $\mathbf{G}(\mathcal{D}) \subset S^n$ and $\mathbf{G}(\mathcal{D}) \cap \mathcal{U} = S^n \cap \mathcal{U}$. Now,

$$\mathbf{DG}(x_1, \ldots, x_n) = \left[\begin{array}{c} I_n \\ \hline \mathbf{v} \end{array}\right],$$

where $\mathbf{v} = \nabla G_{n+1}(x_1, \ldots, x_n)$. Since the first n-rows of $\mathbf{DG}(x_1, \ldots, x_n)$ is the $n \times n$ identity matrix, it has rank n. Thus, S^n is an n-dimensional surface.

Generalizing Example 5.13, we find that a large class of surfaces is provided by graphs of differentiable functions.

Theorem 5.12

Let $\mathcal{D}$ be an open subset of $\mathbb{R}^n$, and let $g : \mathcal{D} \to \mathbb{R}$ be a differentiable mapping. Then the graph of g given by

$$G_g = \{(x_1, \ldots, x_n, x_{n+1}) \,|\, (x_1, \ldots, x_n) \in \mathcal{D}, x_{n+1} = g(x_1, \ldots, x_n)\},$$

is an n-dimensional surface.

A hyperplane in $\mathbb{R}^{n+1}$ is the set of points in $\mathbb{R}^{n+1}$ which satisfies an equation

of the form

$$a_1x_1 + \cdots + a_nx_n + a_{n+1}x_{n+1} = b,$$

where $\mathbf{a} = (a_1, \ldots, a_n, a_{n+1})$ is a nonzero vector in $\mathbb{R}^{n+1}$. By definition, if $\mathbf{u}$ and $\mathbf{v}$ are two points on the plane, then

$$\langle \mathbf{a}, \mathbf{u} - \mathbf{v} \rangle = \mathbf{0}.$$

This shows that $\mathbf{a}$ is a vector normal to the plane.

When $\mathcal{D}$ is an open subset of $\mathbb{R}^n$, and $g : \mathcal{D} \to \mathbb{R}$ is a differentiable mapping, the graph G_g of g is an n-dimensional surface. If $\mathbf{u} = (u_1, \ldots, u_n)$ is a point on $\mathcal{D}$, $(\mathbf{u}, g(\mathbf{u}))$ is a point on G_g, we have seen that the equation of the tangent plane at the point $(\mathbf{u}, g(\mathbf{u}))$ is given by

$$x_{n+1} = f(\mathbf{u}) + \sum_{i=1}^{n} \frac{\partial g}{\partial x_i}(\mathbf{u}, g(\mathbf{u}))(x_i - u_i).$$

Implicit function theorem gives the following.

Theorem 5.13

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^{n+1}$, and let $f : \mathcal{O} \to \mathbb{R}$ be a continuously differentiable function. If $\mathbf{x}_0$ is a point in $\mathcal{O}$ such that $f(\mathbf{x}_0) = 0$ and $\nabla f(\mathbf{x}_0) \neq \mathbf{0}$, then there is neighbourhood U of $\mathbf{x}_0$ contained in $\mathcal{O}$ such that restricted to U, $f(\mathbf{x}) = 0$ is the graph of a continuously differentiable function $g : \mathcal{D} \to \mathbb{R}$, and $\nabla f(\mathbf{x})$ is a vector normal to the tangent plane of the graph at the point $\mathbf{x}$.

Proof

Assume that $\mathbf{x}_0 = (a_1, \ldots, a_n, a_{n+1})$. Since $\nabla f(\mathbf{x}_0) \neq 0$, there is a $1 \leq k \leq n+1$ such that $\dfrac{\partial f}{\partial x_k}(\mathbf{x}_0) \neq 0$. Without loss of generality, assume that $k = n + 1$.

Given a point $\mathbf{x} = (x_1, \ldots, x_n, x_{n+1})$ in $\mathbb{R}^{n+1}$, let $\mathbf{u} = (x_1, \ldots, x_n)$ so that $\mathbf{x} = (\mathbf{u}, x_{n+1})$. By the implicit function theorem, there is a neighbourhood $\mathcal{D}$ of $\mathbf{u}_0 = (a_1, \ldots, a_n)$, an $r > 0$, and a continuously differentiable function $g : \mathcal{D} \to \mathbb{R}$ such that if $U = \mathcal{D} \times (a_{n+1} - r, a_{n+1} + r)$, $(\mathbf{u}, u_{n+1}) \in U$, then $f(\mathbf{u}, u_{n+1}) = 0$ if and only if $u_{n+1} = g(\mathbf{u})$. In other words, in the neighbourhood U of $\mathbf{x}_0 = (\mathbf{u}_0, a_{n+1})$, $f(\mathbf{u}, u_{n+1}) = 0$ if and only if $(\mathbf{u}, u_{n+1})$ is a point on the graph of the function g. The equation of the tangent plane at the point $(\mathbf{u}, u_{n+1})$ is

$$x_{n+1} - u_{n+1} = \sum_{i=1}^{n} \frac{\partial g}{\partial x_i}(\mathbf{u})(x_i - u_i).$$

By chain rule,

$$\frac{\partial g}{\partial x_i}(\mathbf{u}) = -\frac{\dfrac{\partial f}{\partial x_i}(\mathbf{u}, u_{n+1})}{\dfrac{\partial f}{\partial x_{n+1}}(\mathbf{u}, u_{n+1})}.$$

Hence, the equation of the tangent plane can be rewritten as

$$\sum_{i=1}^{n+1}(x_i - u_i)\frac{\partial f}{\partial x_i}(\mathbf{u}, u_{n+1}) = 0.$$

This shows that $\nabla f(\mathbf{u}, u_{n+1})$ is a vector normal to the tangent plane.

Example 5.14

Find the equation of the tangent plane to the surface $x^2 + 4y^2 + 9z^2 = 36$ at the point $(6, 1, -1)$.

Solution

Let $f(x, y, z) = x^2 + 4y^2 + 9z^2$. Then $\nabla f(x, y, z) = (2x, 8y, 18z)$. It follows that $\nabla f(6, 1, -1) = 2(6, 4, -9)$. Hence, the equation of the tangent plane to the surface at $(6, 1, -1)$ is

$$6x + 4y - 9z = 36 + 4 + 9 = 49.$$

Exercises 5.3

Question 1

Consider the equation

$$4yz^2 + 3xz^3 - 11xyz = 14.$$

Show that in a neighbourhood of $(-1, 1, 2)$, this equation defines z as a function of (x, y). If this function is denoted as $z = g(x, y)$, find $\nabla g(-1, 1)$.

Question 2

Consider the system of equations

$$\begin{aligned} 2xu^2 + vyz + 3uv &= 2 \\ 5x + 7yzu - v^2 &= 1 \end{aligned}$$

(a) Show that when $(x, y, z) = (-1, 1, 1)$, $(u, v) = (1, 1)$ is a solution of this system.

(b) Show that there are neighbourhoods U and V of $(-1, 1, 1)$ and $(1, 1)$, and a continuously differentiable function $\mathbf{G} : U \to \mathbb{R}^2$ such that, if $(x, y, z, u, v) \in U \times V$, then (x, y, z, u, v) is a solution of the system of equations above if and only if $u = G_1(x, y, z)$ and $v = G_2(x, y, z)$.

(c) Find the values of $\dfrac{\partial G_1}{\partial x}(-1, 1, 1)$, $\dfrac{\partial G_2}{\partial x}(-1, 1, 1)$ and $\dfrac{\partial G_2}{\partial z}(-1, 1, 1)$.

Question 3

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^{2n}$, and let $\mathbf{F} : \mathcal{O} \to \mathbb{R}^n$ be a continuously differentiable function. Assume that $\mathbf{x}_0$ and $\mathbf{y}_0$ are points in $\mathbb{R}^n$ such that $(\mathbf{x}_0, \mathbf{y}_0)$ is a point in $\mathcal{O}$, $\mathbf{F}(\mathbf{x}_0, \mathbf{y}_0) = \mathbf{0}$, and $\mathbf{D}_{\mathbf{x}}\mathbf{F}(\mathbf{x}_0, \mathbf{y}_0)$ and $\mathbf{D}_{\mathbf{y}}\mathbf{F}(\mathbf{x}_0, \mathbf{y}_0)$ are invertible. Show that there exist neighbourhoods U and V of $\mathbf{x}_0$ and $\mathbf{y}_0$, and a continuously differentiable bijective function $\mathbf{G} : U \to V$ such that, if $(\mathbf{x}, \mathbf{y})$ is in $U \times V$, $\mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ if and only if $\mathbf{y} = \mathbf{G}(\mathbf{x})$.

5.4 Extrema Problems and the Method of Lagrange Multipliers

Optimization problems are very important in our daily life and in mathematical sciences. Given a function $f : \mathfrak{D} \to \mathbb{R}$, we would like to know whether it has a maximum value or a minimum value. In Chapter 3, we have dicusssed the extreme value theorem, which asserts that a continuous function that is defined on a compact set must have maximum and minimum values. In Chapter 4, we showed that if a function $f : \mathfrak{D} \to \mathbb{R}$ has (local) extremum at an interior point $\mathbf{x}_0$ of its domain $\mathfrak{D}$ and it is differentiable at $\mathbf{x}_0$, then $\mathbf{x}_0$ must be a stationary point. Namely, $\nabla f(\mathbf{x}_0) = \mathbf{0}$.

Combining these various results, we can formulate a strategy for solving a special type of optimization problems. Let us first consider the following example.

Example 5.15

Let

$$K = \left\{(x, y) \,|\, x^2 + 4y^2 \leq 100\right\},$$

and let $f : K \to \mathbb{R}$ be the function defined as

$$f(x, y) = x^2 + y^2.$$

Find the maximum and minimum values of $f : K \to \mathbb{R}$, and the points where these values appear.

Solution

Let $g : \mathbb{R}^2 \to \mathbb{R}$ be the function defined as $g(x, y) = x^2 + 4y^2 - 100$. It is a polynomial function. Hence, it is continuous. Since $K = g^{-1}((-\infty, 0])$ and $(-\infty, 0]$ is closed in $\mathbb{R}$, K is a closed set. By a previous exercise,

$$\mathcal{O} = \operatorname{int} K = \left\{(x, y) \,|\, x^2 + 4y^2 < 100\right\}$$

and

$$\mathcal{C} = \operatorname{bd} K = \left\{(x, y) \,|\, x^2 + 4y^2 = 100\right\}.$$

For any $(x, y) \in K$, $\|(x, y)\|^2 = x^2 + y^2 \leq x^2 + 4y^2 \leq 100$. Therefore, K is bounded. Since K is closed and bounded, and the function $f : K \to \mathbb{R}$, $f(x) = x^2 + y^2$ is continuous, extreme value theorem says that f has maximum and minimum values. These values appear either in $\mathcal{O}$ or on $\mathcal{C}$.
Since $f : \mathcal{O} \to \mathbb{R}$ is differentiable, if (x_0, y_0) is an extremizer of $f : \mathcal{O} \to \mathbb{R}$, we must have $\nabla f(x_0, y_0) = (0, 0)$, which gives $(x_0, y_0) = (0, 0)$.
The other candidates of extremizers are on $\mathcal{C}$. Therefore, we need to find the maximum and minimum values of $f(x, y) = x^2 + y^2$ subject to the constraint $x^2 + 4y^2 = 100$. From $x^2 + 4y^2 = 100$, we find that $x^2 = 100 - 4y^2$, and y can only take values in the interval $[-5, 5]$. Hence, we want to find the maximum and minimum values of $h : [-5, 5] \to \mathbb{R}$,

$$h(y) = 100 - 4y^2 + y^2 = 100 - 3y^2.$$

When $y = 0$, h has maximum value 100, and when $y = \pm 5$, it has minimum value $100 - 3 \times 25 = 25$. Notice that when $y = 0$, $x = \pm 10$; while when $y = \pm 5$, $x = 0$.
Hence, we have five candidates for the extremizers of f. Namely, $\mathbf{u}_1 = (0, 0)$, $\mathbf{u}_2 = (10, 0)$, $\mathbf{u}_3 = (-10, 0)$, $\mathbf{u}_4 = (0, 5)$ and $\mathbf{u}_5 = (0, -5)$. The function values at these 5 points are

$$f(\mathbf{u}_1) = 0, \quad f(\mathbf{u}_2) = f(\mathbf{u}_3) = 100, \quad f(\mathbf{u}_4) = f(\mathbf{u}_5) = 25.$$

Therefore, the minimum value of $f : K \to \mathbb{R}$ is 0, and the maximum value is 100. The minimum value appears at the point $(0, 0) \in \operatorname{int} K$, while the maximum value appears at $(\pm 10, 0) \in \operatorname{bd} K$.

Example 5.15 gives a typical scenario of the optimization problems that we want to study in this section.

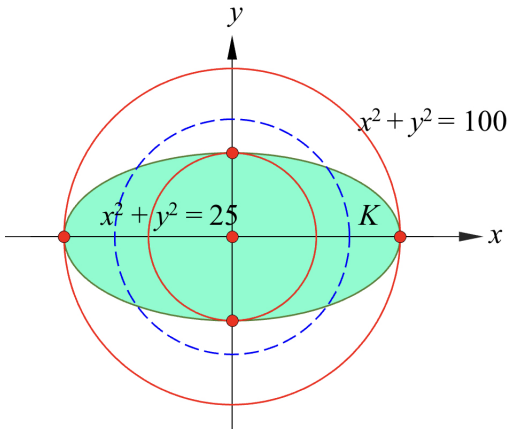


Figure 5.7: The extreme values of $f(x,y) = x^2 + y^2$ on the sets $K = \{(x,y) \,|\, x^2 + 4y^2 \le 100\}$ and $\mathcal{C} = \{(x,y) \,|\, x^2 + 4y^2 = 100\}$.

Optimization Problem

Let K be a compact subset of $\mathbb{R}^n$ with interior $\mathcal{O}$, and let $f : K \to \mathbb{R}$ be a function continuous on K, differentiable on $\mathcal{O}$. We want to find the maximum and minimum values of $f : K \to \mathbb{R}$.

(i) By the extreme value theorem, $f : K \to \mathbb{R}$ has maximum and minimum values.

(ii) Since K is closed, K is a disjoint union of its interior $\mathcal{O}$ and its boundary $\mathcal{C}$. Since $\mathcal{C}$ is a subset of K, it is bounded. On the other hand, being the boundary of a set, $\mathcal{C}$ is closed. Therefore, $\mathcal{C}$ is compact.

(iii) The extreme values of f can appear in $\mathcal{O}$ or on $\mathcal{C}$.

(iv) If $\mathbf{x}_0$ is an extremizer of $f : K \to \mathbb{R}$ and it is in $\mathcal{O}$, we must have $\nabla f(\mathbf{x}_0) = \mathbf{0}$. Namely, $\mathbf{x}_0$ is a stationary point of $f : \mathcal{O} \to \mathbb{R}$.

(v) If $\mathbf{x}_0$ is an extremizer of $f : K \to \mathbb{R}$ and it is not in $\mathcal{O}$, it is an extremizer of $f : \mathcal{C} \to \mathbb{R}$.

(vi) Since $\mathcal{C}$ is compact, $f : \mathcal{C} \to \mathbb{R}$ has maximum and minimum values.

Therefore, the steps to find the maximum and minimum values of $f : K \to \mathbb{R}$ are as follows.

Step 1 Find the stationary points of $f : \mathcal{O} \to \mathbb{R}$.

Step 2 Find the extremizers of $f : \mathcal{C} \to \mathbb{R}$.

Step 3 Compare the values of f at the stationary points of $f : \mathcal{O} \to \mathbb{R}$ and the extremizers of $f : \mathcal{C} \to \mathbb{R}$ to determine the extreme values of $f : K \to \mathbb{R}$.

Of particular interest is when the boundary of K can be expressed as $g(\mathbf{x}) = 0$, where $g : \mathcal{D} \to \mathbb{R}$ is a continuously differentiable function defined on an open subset $\mathcal{D}$ of $\mathbb{R}^n$. If f is also defined and differentiable on $\mathcal{D}$, the problem of finding the extreme values of $f : \mathcal{C} \to \mathbb{R}$ becomes finding the extreme values of $f : \mathcal{D} \to \mathbb{R}$ subject to the constraint $g(\mathbf{x}) = 0$. In Example 5.15, we have used $g(\mathbf{x}) = 0$ to solve one of the variables in terms of the others and substitute into f to transform the optimization problem to a problem with fewer variables. However, this strategy can be quite complicated because it is often not possible to solve one variable in terms of the others explicitly from the constraint $g(\mathbf{x}) = 0$. The method of Lagrange multipliers provides a way to solve constraint optimization problems without having to explicitly solve some variables in terms of the others. The validity of this method is justified by the implicit function theorem.

Theorem 5.14 The Method of Lagrange Multiplier (One Constraint)

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^{n+1}$ and let $f : \mathcal{O} \to \mathbb{R}$ and $g : \mathcal{O} \to \mathbb{R}$ be continuously differentiable functions defined on $\mathcal{O}$. Consider the subset of $\mathcal{O}$ defined as

$$\mathcal{C} = \{\mathbf{x} \in \mathcal{O} \,|\, g(\mathbf{x}) = 0\}.$$

If $\mathbf{x}_0$ is an extremizer of the function $f : \mathcal{C} \to \mathbb{R}$ and $\nabla g(\mathbf{x}_0) \neq \mathbf{0}$, then there is a constant λ, known as the Lagrange multiplier, such that

$$\nabla f(\mathbf{x}_0) = \lambda \nabla g(\mathbf{x}_0).$$

Proof

Without loss of generality, assume that $\mathbf{x}_0$ is a maximizer of $f : \mathcal{C} \to \mathbb{R}$. Namely,

$$f(\mathbf{x}) \leq f(\mathbf{x}_0) \qquad \text{for all } \mathbf{x} \in \mathcal{C}. \tag{5.9}$$

Given that $\nabla g(\mathbf{x}_0) \neq 0$, there exists a $1 \leq k \leq n+1$ such that $\dfrac{\partial g}{\partial x_k}(\mathbf{x}_0) \neq 0$. Without loss of generality, assume that $k = n+1$. Let $\mathbf{x}_0 = (a_1, \ldots, a_n, a_{n+1})$. Given a point $\mathbf{x} = (x_1, \ldots, x_n, x_{n+1})$ in $\mathbb{R}^{n+1}$, let $\mathbf{u} = (x_1, \ldots, x_n)$ so that $\mathbf{x} = (\mathbf{u}, x_{n+1})$. By implicit function theorem, there is a neighbourhood $\mathcal{D}$ of $\mathbf{u}_0 = (a_1, \ldots, a_n)$, an $r > 0$, and a continuously differentiable function $h : \mathcal{D} \to \mathbb{R}$ such that for $(\mathbf{u}, x_{n+1}) \in \mathcal{D} \times (a_{n+1} - r, a_{n+1} + r)$, $g(\mathbf{u}, x_{n+1}) = 0$ if and only if $x_{n+1} = h(\mathbf{u})$. Consider the function $F : \mathcal{D} \to \mathbb{R}$ defined as

$$F(\mathbf{u}) = f(\mathbf{u}, h(\mathbf{u})).$$

By (5.9), we find that

$$F(\mathbf{u}_0) \geq F(\mathbf{u}) \qquad \text{for all } \mathbf{u} \in \mathcal{D}.$$

In other words, $\mathbf{u}_0$ is a maximizer of the function $F : \mathcal{D} \to \mathbb{R}$. Since $\mathbf{u}_0$ is an interior point of $\mathcal{D}$ and $F : \mathcal{D} \to \mathbb{R}$ is continuously differentiable, $\nabla F(\mathbf{u}_0) = 0$. Since $F(\mathbf{u}) = f(\mathbf{u}, h(\mathbf{u}))$, we find that for $1 \leq i \leq n$,

$$\frac{\partial F}{\partial x_i}(\mathbf{u}_0) = \frac{\partial f}{\partial x_i}(\mathbf{u}_0, a_{n+1}) + \frac{\partial f}{\partial x_{n+1}}(\mathbf{u}_0, a_{n+1}) \frac{\partial h}{\partial x_i}(\mathbf{u}_0) = 0. \tag{5.10}$$

On the other hand, applying chain rule to $g(\mathbf{u}, h(\mathbf{u})) = 0$ and set $\mathbf{u} = \mathbf{u}_0$, we find that

$$\frac{\partial g}{\partial x_i}(\mathbf{u}_0, a_{n+1}) + \frac{\partial g}{\partial x_{n+1}}(\mathbf{u}_0, a_{n+1}) \frac{\partial h}{\partial x_i}(\mathbf{u}_0) = 0 \qquad \text{for } 1 \leq i \leq n. \tag{5.11}$$

By assumption, $\dfrac{\partial g}{\partial x_{n+1}}(\mathbf{x}_0) \neq 0$. Let

$$\lambda = \frac{\dfrac{\partial f}{\partial x_{n+1}}(\mathbf{x}_0)}{\dfrac{\partial g}{\partial x_{n+1}}(\mathbf{x}_0)}.$$

Then

$$\frac{\partial f}{\partial x_{n+1}}(\mathbf{x}_0) = \lambda \frac{\partial g}{\partial x_{n+1}}(\mathbf{x}_0). \tag{5.12}$$

Eqs. (5.10) and (5.11) show that for $1 \leq i \leq n$,

$$\frac{\partial f}{\partial x_i}(\mathbf{x}_0) = -\lambda \frac{\partial g}{\partial x_{n+1}}(\mathbf{x}_0) \frac{\partial h}{\partial x_i}(\mathbf{u}_0) = \lambda \frac{\partial g}{\partial x_i}(\mathbf{x}_0). \tag{5.13}$$

Eqs. (5.12) and (5.13) together imply that

$$\nabla f(\mathbf{x}_0) = \lambda \nabla g(\mathbf{x}_0).$$

This completes the proof of the theorem.

Remark 5.5

Theorem 5.14 says that if $\mathbf{x}_0$ is an extremizer of the constraint optimization problem $\max / \min f(\mathbf{x})$ subject to $g(\mathbf{x}) = 0$, then the gradient of f at $\mathbf{x}_0$ should be parallel to the gradient of g at $\mathbf{x}_0$ if the latter is nonzero. One can refer to Figure 5.7 for an illustration. Recall that the gradient of f gives the direction where f changes most rapidly, while the gradient of g here represents the normal vector to the curve $g(\mathbf{x}) = 0$.

Using the method of Lagrange multiplier, there are $n + 2$ variables $x_1, \ldots, x_{n+1}$ and λ to be solved. The equation $\nabla f(\mathbf{x}) = \lambda \nabla g(\mathbf{x})$ gives $n + 1$ equations, while the equation $g(\mathbf{x}) = 0$ gives one. Therefore, we need to solve $n + 2$ variables from $n + 2$ equations.

Example 5.16

Let us solve the constraint optimization problem that appears in Example 5.15 using the Lagrange multiplier method. Let $f : \mathbb{R}^2 \to \mathbb{R}$ and $g : \mathbb{R}^2 \to \mathbb{R}$ be respectively the functions $f(x, y) = x^2 + y^2$ and $g(x, y) = x^2 + 4y^2 - 100$. They are both continuously differentiable. We want to find the maximum and minimum values of the function $f(x, y)$ subject to the constraint $g(x, y) = 0$. Notice that $\nabla g(x, y) = (2x, 8y)$ is the zero vector if and only if $(x, y) = (0, 0)$, but $(0, 0)$ is not on the curve $g(x, y) = 0$. Hence, for any (x, y) satisfying $g(x, y) = 0$, $\nabla g(x, y) \neq \mathbf{0}$.

By the method of Lagrange multiplier, we need to find (x, y) satisfying

$$\nabla f(x, y) = \lambda \nabla g(x, y) \quad \text{and} \quad g(x, y) = 0.$$

Therefore,

$$2x = 2\lambda x, \qquad 2y = 8\lambda y.$$

This gives

$$x(1 - \lambda) = 0, \quad y(1 - 4\lambda) = 0.$$

The first equation says that either $x = 0$ or $\lambda = 1$.
If $x = 0$, from $x^2 + 4y^2 = 100$, we must have $y = \pm 5$.
If $\lambda = 1$, then $y(1 - 4\lambda) = 0$ implies that $y = 0$. From $x^2 + 4y^2 = 100$, we then obtain $x = \pm 10$.
Hence, we find that the candidates for the extremizers are $(\pm 10, 0)$ and $(0, \pm 5)$. Since $f(\pm 10, 0) = 100$ and $f(0, \pm 5) = 25$, we conclude that subject to $x^2 + 4y^2 = 100$, the maximum value of $f(x, y) = x^2 + y^2$ is 100, and the minimum value of $f(x, y) = x^2 + y^2$ is 25.

Example 5.17

Use the Lagrange multiplier method to find the maximum and minimum values of the function $f(x, y, z) = 8x + 24y + 27z$ on the set

$$S = \left\{(x, y, z) \,|\, x^2 + 4y^2 + 9z^2 = 289\right\},$$

and the points where each of them appears.

Solution

Let $g : \mathbb{R}^3 \to \mathbb{R}$ be the function

$$g(x, y, z) = x^2 + 4y^2 + 9z^2 - 289.$$

The functions $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x, y, z) = 8x + 24y + 27z$ and $g : \mathbb{R}^3 \to \mathbb{R}$ are both continuously differentiable.

Notice that $\nabla g(x,y,z) = (2x, 8y, 18z) = \mathbf{0}$ if and only if $(x,y,z) = \mathbf{0}$, and $\mathbf{0}$ does not lie on S. By Lagrange multiplier method, to find the maximum and minimum values of $f : S \to \mathbb{R}$, we need to solve the equations

$$\nabla f(x,y,z) = \lambda \nabla g(x,y,z) \quad \text{and} \quad g(x,y,z) = 0.$$

These give

$$8 = 2\lambda x, \quad 24 = 8\lambda y, \quad 27 = 18\lambda z$$
$$x^2 + 4y^2 + 9z^2 = 289.$$

To satisfy the first three equations, none of the λ, x, y and z can be zero. We find that

$$x = \frac{4}{\lambda}, \quad y = \frac{3}{\lambda}, \quad z = \frac{3}{2\lambda}.$$

Substitute into the last equation, we have

$$\frac{64 + 144 + 81}{4\lambda^2} = 289.$$

This gives $4\lambda^2 = 1$. Hence, $\lambda = \pm\frac{1}{2}$. When $\lambda = \frac{1}{2}$, $(x,y,z) = (8,6,3)$. When $\lambda = -\frac{1}{2}$, $(x,y,z) = (-8,-6,-3)$. These are the two candidates for the extremizers of $f : S \to \mathbb{R}$.

Since $f(8,6,3) = 289$ and $f(-8,-6,-3) = -289$, we find that the maximum and minimum values of $f : S \to \mathbb{R}$ are 289 and -289 respectively, and the maximum value appear at $(8,6,3)$, the minimum value appear at $(-8,-6,-3)$.

Now we consider more general constraint optimization problems which can have more than one constraints.

Theorem 5.15 The Method of Lagrange Multiplier (General)

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^{m+n}$ and let $f : \mathcal{O} \to \mathbb{R}$ and $\mathbf{G} : \mathcal{O} \to \mathbb{R}^m$ be continuously differentiable functions defined on $\mathcal{O}$. Consider the subset of $\mathcal{O}$ defined as

$$\mathcal{C} = \{\mathbf{x} \in \mathcal{O} \,|\, \mathbf{G}(\mathbf{x}) = \mathbf{0}\}.$$

If $\mathbf{x}_0$ is an extremizer of the function $f : \mathcal{C} \to \mathbb{R}$ and the matrix $\mathrm{DG}(\mathbf{x}_0)$ has (maximal) rank m, then there are constants $\lambda_1, \ldots, \lambda_m$, known as the Lagrange multipliers, such that

$$\nabla f(\mathbf{x}_0) = \sum_{i=1}^{m} \lambda_i \nabla G_i(\mathbf{x}_0).$$

Proof

Without loss of generality, assume that $\mathbf{x}_0$ is a maximizer of $f : \mathcal{C} \to \mathbb{R}$. Namely,

$$f(\mathbf{x}) \leq f(\mathbf{x}_0) \qquad \text{for all } \mathbf{x} \in \mathcal{C}. \tag{5.14}$$

Given that the matrix $\mathbf{DG}(\mathbf{x}_0)$ has rank m, m of the column vectors are linearly independent. Without loss of generality, assume that the column vectors in the last m columns are linearly independent. Write a point $\mathbf{x}$ in $\mathbb{R}^{m+n}$ as $\mathbf{x} = (\mathbf{u}, \mathbf{v})$, where $\mathbf{u} = (u_1, \ldots, u_n)$ is in $\mathbb{R}^n$ and $\mathbf{v} = (v_1, \ldots, v_m)$ is in $\mathbb{R}^m$. By our assumption, $\mathbf{D}_{\mathbf{v}}\mathbf{G}(\mathbf{u}_0, \mathbf{v}_0)$ is invertible. By implicit function theorem, there is a neighbourhood $\mathcal{D}$ of $\mathbf{u}_0$, a neighbourhood $\mathcal{V}$ of $\mathbf{v}_0$, and a continuously differentiable function $\mathbf{H} : \mathcal{D} \to \mathbb{R}^m$ such that for $(\mathbf{u}, \mathbf{v}) \in \mathcal{D} \times \mathcal{V}$, $\mathbf{G}(\mathbf{u}, \mathbf{v}) = \mathbf{0}$ if and only if $\mathbf{v} = \mathbf{H}(\mathbf{u})$. Consider the function $F : \mathcal{D} \to \mathbb{R}$ defined as

$$F(\mathbf{u}) = f(\mathbf{u}, \mathbf{H}(\mathbf{u})).$$

By (5.14), we find that

$$F(\mathbf{u}_0) \geq F(\mathbf{u}) \qquad \text{for all } \mathbf{u} \in \mathcal{D}.$$

In other words, $\mathbf{u}_0$ is a maximizer of the function $F : \mathcal{D} \to \mathbb{R}$. Since $\mathbf{u}_0$ is an interior point of $\mathcal{D}$ and $F : \mathcal{D} \to \mathbb{R}$ is continuously differentiable, $\nabla F(\mathbf{u}_0) = 0$. Since $F(\mathbf{u}) = f(\mathbf{u}, \mathbf{H}(\mathbf{u}))$, we find that

$$\nabla F(\mathbf{u}_0) = D_{\mathbf{u}} f(\mathbf{u}_0, \mathbf{v}_0) + D_{\mathbf{v}} f(\mathbf{u}_0, \mathbf{v}_0)\mathbf{D}\mathbf{H}(\mathbf{u}_0) = \mathbf{0}. \tag{5.15}$$

On the other hand, applying chain rule to $\mathbf{G}(\mathbf{u}, \mathbf{H}(\mathbf{u})) = \mathbf{0}$ and set $\mathbf{u} = \mathbf{u}_0$, we find that

$$\mathbf{D}_{\mathbf{u}}\mathbf{G}(\mathbf{u}_0, \mathbf{v}_0) + \mathbf{D}_{\mathbf{v}}\mathbf{G}(\mathbf{u}_0, \mathbf{v}_0)\mathbf{D}\mathbf{H}(\mathbf{u}_0) = \mathbf{0}. \tag{5.16}$$

Take

$$\begin{bmatrix} \lambda_1 & \lambda_2 & \cdots & \lambda_m \end{bmatrix} = \boldsymbol{\lambda} = D_{\mathbf{v}} f(\mathbf{x}_0)\mathbf{D}_{\mathbf{v}}\mathbf{G}(\mathbf{x}_0)^{-1}.$$

Then

$$D_{\mathbf{v}} f(\mathbf{x}_0) = \boldsymbol{\lambda} D_{\mathbf{v}}\mathbf{G}(\mathbf{x}_0). \tag{5.17}$$

Eqs. (5.15) and (5.16) show that

$$D_{\mathbf{u}} f(\mathbf{x}_0) = -\boldsymbol{\lambda}\mathbf{D}_{\mathbf{v}}\mathbf{G}(\mathbf{x}_0)\mathbf{D}\mathbf{H}(\mathbf{u}_0) = \boldsymbol{\lambda}\mathbf{D}_{\mathbf{u}}\mathbf{G}(\mathbf{x}_0). \tag{5.18}$$

Eqs. (5.17) and (5.18) together imply that

$$\nabla f(\mathbf{x}_0) = \boldsymbol{\lambda}\mathbf{D}\mathbf{G}(\mathbf{x}_0) = \sum_{i=1}^{m} \lambda_i \nabla G_i(\mathbf{x}_0).$$

This completes the proof of the theorem.

In the general constraint optimization problem proposed in Theorem 5.15, there are $n + 2m$ variables $u_1, \ldots, u_n$, $v_1, \ldots, v_m$ and $\lambda_1, \ldots, \lambda_m$ to be solved. The components of

$$\nabla f(\mathbf{x}) = \sum_{i=1}^{m} \lambda_i \nabla G_i(\mathbf{x})$$

give $n + m$ equations, while the components of $\mathbf{G}(\mathbf{x}) = \mathbf{0}$ give m equations. Hence, we have to solve $n + 2m$ variables from $n + 2m$ equations. Let us look at an example.

Example 5.18

Let K be the subset of $\mathbb{R}^3$ given by

$$K = \left\{(x, y, z) \,|\, x^2 + y^2 \leq 4, x + y + z = 1\right\}.$$

Find the maximum and minimum values of the function $f : K \to \mathbb{R}$, $f(x, y, z) = x + 3y + z$.

Solution

Notice that K is the intersection of the two closed sets $K_1 = \{(x, y, z) \,|\, x^2 + y^2 \leq 4\}$ and $K_2 = \{(x, y, z) \,|\, x + y + z = 1\}$. Hence, K is a closed set. If (x, y, z) is in K, $x^2 + y^2 \leq 4$. Thus, $|x| \leq 2$, $|y| \leq 2$ and hence $|z| \leq 1 + |x| + |y| \leq 5$. This shows that K is bounded. Since K is closed and bounded, $f : K \to \mathbb{R}$ is continuous, $f : K \to \mathbb{R}$ has maximum and minimum values.

Let

$$D = \left\{(x, y, z) \,|\, x^2 + y^2 < 4, x + y + z = 1\right\},$$

$$C = \left\{(x, y, z) \,|\, x^2 + y^2 = 4, x + y + z = 1\right\}.$$

Then $K = C \cup D$. We can consider the extremizers of $f : D \to \mathbb{R}$ and $f : C \to \mathbb{R}$ separately.

To find the extremizers of $f : D \to \mathbb{R}$, we can regard this as a constraint optimization problem where we want to find the extreme values of $f : \mathcal{O} \to \mathbb{R}$, $f(x, y, z) = x + 3y + z$ on

$$\mathcal{O} = \left\{(x, y, z) \,|\, x^2 + y^2 < 4\right\},$$

subject to the constraint $g(x, y, z) = 0$, where $g : \mathcal{O} \to \mathbb{R}$ is the function $g(x, y, z) = x + y + z - 1$. Now $\nabla g(x, y, z) = (1, 1, 1) \neq \mathbf{0}$. Hence, at an extremizer, we must have $\nabla f(x, y, z) = \lambda g(x, y, z)$, which gives

$$(1, 3, 1) = \lambda(1, 1, 1).$$

This says that the two vectors $(1,3,1)$ and $(1,1,1)$ must be parallel, which is a contradiction. Hence, $f : \mathcal{O} \to \mathbb{R}$ does not have extremizers.
Now, to find the extremizers of $f : C \to \mathbb{R}$, we can consider it as finding the extreme values of $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x,y,z) = x + 3y + z$, subject to $\mathbf{G}(x,y,z) = 0$, where

$$\mathbf{G}(x,y,z) = (x^2 + y^2 - 4, x + y + z - 1).$$

Now

$$\mathbf{DG}(x,y,z) = \begin{bmatrix} 2x & 2y & 0 \\ 1 & 1 & 1 \end{bmatrix}.$$

This matrix has rank less than 2 if and only if $(2x, 2y, 0)$ is parallel to $(1,1,1)$, which gives $x = y = z = 0$. But the point $(x,y,z) = (0,0,0)$ is not on C. Therefore, $\mathbf{DG}(x,y,z)$ has maximal rank for every $(x,y,z) \in C$. Using the Lagrange multiplier method, to solve for the extremizer of $f : C \to \mathbb{R}$, we need to solve the system

$$\nabla f(x,y,z) = \lambda \nabla G_1(x,y,z) + \mu G_2(x,y,z), \quad \mathbf{G}(x,y,z) = \mathbf{0}.$$

These gives

$$1 = 2\lambda x + \mu, \quad 3 = 2\lambda y + \mu, \quad 1 = \mu,$$
$$x^2 + y^2 = 4, \quad x + y + z = 1.$$

From $\mu = 1$, we have $2\lambda x = 0$ and $2\lambda y = 2$. The latter implies that $\lambda \neq 0$. Hence, we must have $x = 0$. Then $x^2 + y^2 = 4$ gives $y = \pm 2$. When $(x,y) = (0,2)$, $z = -1$. When $(x,y) = (0,-2)$, $z = 3$. Hence, we only have two candidates for extremizers, which are $(0,2,-1)$ and $(0,-2,3)$. Since

$$f(0,2,-1) = 5, \quad f(0,-2,3) = -3,$$

we find that $f : K \to \mathbb{R}$ has maximum value 5 at the point $(0,2,-1)$, and minimum value -3 at the point $(0,-2,3)$.

Exercises 5.4

Question 1

Find the extreme values of the function $f(x, y, z) = 4x^2 + y^2 + yz + z^2$ on the set

$$S = \left\{(x, y, z) \,|\, 2x^2 + y^2 + z^2 \leq 8\right\}.$$

Question 2

Find the point in the set

$$S = \left\{(x, y) \,|\, 4x^2 + y^2 \leq 36, x^2 + 4y^2 \geq 4\right\}$$

that is closest to and farthest from the point $(1, 0)$.

Question 3

Use the Lagrange multiplier method to find the maximum and minimum values of the function $f(x, y, z) = x + 2y - z$ on the set

$$S = \left\{(x, y, z) \,|\, x^2 + y^2 + 4z^2 \leq 84\right\},$$

and the points where each of them appears.

Question 4

Find the extreme values of the function $f(x, y, z) = x$ on the set

$$S = \left\{(x, y, z) \,|\, x^2 = y^2 + z^2, 7x + 3y + 4z = 60\right\}.$$

Question 5

Let K be the subset of $\mathbb{R}^3$ given by

$$K = \left\{(x, y, z) \,|\, 4x^2 + z^2 \leq 68, y + z = 12\right\}.$$

Find the maximum and minimum values of the function $f : K \to \mathbb{R}$, $f(x, y, z) = x + 2y$.

Question 6

Let A be an $n \times n$ symmetric matrix, and let $Q_A : \mathbb{R}^n \to \mathbb{R}$ be the quadratic form $Q_A(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ defined by A. Show that the minimum and maximum values of $Q_A : S^{n-1} \to \mathbb{R}$ on the unit sphere S^{n-1} are the smallest and largest eigenvalues of A.

Chapter 6

Multiple Integrals

For a single variable functions, we have discussed the Riemann integrability of a function $f : [a, b] \to \mathbb{R}$ defined on a compact interval $[a, b]$. In this chapter, we consider the theory of Riemann integrals for multivariable functions. For a function $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ that takes values in $\mathbb{R}^m$ with $m \geq 2$, we define the integral componentwise. Namely, we say that the function $\mathbf{F} : \mathfrak{D} \to \mathbb{R}^m$ is Riemann integrable if and only if each of the component functions $F_j : \mathfrak{D} \to \mathbb{R}$, $1 \leq j \leq m$ is Riemann integrable, and we define

$$\int_{\mathfrak{D}} \mathbf{F} = \left(\int_{\mathfrak{D}} F_1, \int_{\mathfrak{D}} F_2, \ldots, \int_{\mathfrak{D}} F_m \right).$$

Thus, in this chapter, we will only discuss the theory of integration for functions $f : \mathfrak{D} \to \mathbb{R}$ that take values in $\mathbb{R}$.

A direct generalization of a compact interval $[a, b]$ to $\mathbb{R}^n$ is a product of compact intervals $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$, which is a closed rectangle. In this chapter, when we say $\mathbf{I}$ is a rectangle, it means $\mathbf{I}$ can be written as $\prod_{i=1}^{n} [a_i, b_i]$ with $a_i < b_i$ for all $1 \leq i \leq n$. The edges of $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$ are $[a_1, b_1]$, $[a_2, b_2]$, …, $[a_n, b_n]$.

We first discuss the integration theory of functions defined on closed rectangles of the form $\prod_{i=1}^{n} [a_i, b_i]$. For applications, we need to consider functions defined on other subsets $\mathfrak{D}$ of $\mathbb{R}^n$.

One of the most useful theoretical tools for evaluating single integrals is the fundamental theorem of calculus. To apply this tool for multiple integrals, we need to consider iterated integrals. Another useful tool is the change of variables formula. For multivariable functions, the change of variables theorem is much more complicated. Nevertheless, we will discuss these in this chapter.

6.1 Riemann Integrals

In this section, we define the Riemann integral of a function $f : \mathfrak{D} \to \mathbb{R}$ defined on a subset $\mathfrak{D}$ of $\mathbb{R}^n$. We first consider the case where $\mathfrak{D} = \prod_{i-1}^{n}[a_i, b_i]$.

Let us first consider partitions. We say that $P = \{x_0, x_1, \ldots, x_k\}$ is a partition of the interval $[a, b]$ if $a = x_0 < x_1 < \cdots < x_{k-1} < x_k = b$. It divides $[a, b]$ into k subintervals $J_1, \ldots, J_k$, where $J_i = [x_{i-1}, x_i]$.

Definition 6.1 Partitions

A partition $\mathbf{P}$ of a closed rectangle $\mathbf{I} = \prod_{i=1}^{n}[a_i, b_i]$ is achieved by having a partition P_i of $[a_i, b_i]$ for each $1 \leq i \leq n$. We write $\mathbf{P} = (P_1, P_2, \ldots, P_n)$ for such a partition. The partition $\mathbf{P}$ divides the rectangle $\mathbf{I}$ into a collection $\mathcal{J}_\mathbf{P}$ of rectangles, any two of which have disjoint interiors. A closed rectangle $\mathbf{J}$ in $\mathcal{J}_\mathbf{P}$ can be written as

$$\mathbf{J} = J_1 \times J_2 \times \cdots \times J_n,$$

where J_i, $1 \leq i \leq n$ is a subinterval in the partition P_i.

If the partition P_i divides $[a_i, b_i]$ into k_i subintervals, then the partition $\mathbf{P} = (P_1, \ldots, P_n)$ divides the rectangle $\mathbf{I} = \prod_{i=1}^{n}[a_i, b_i]$ into $|\mathcal{J}_\mathbf{P}| = k_1 k_2 \cdots k_n$ rectangles.

Example 6.1

Consider the rectangle $\mathbf{I} = [-2, 9] \times [1, 6]$. Let $P_1 = \{-2, 0, 4, 9\}$ and $P_2 = \{1, 3, 6\}$. The partition P_1 divides the interval $I_1 = [-2, 9]$ into the three subintervals $[-2, 0]$, $[0, 4]$ and $[4, 9]$. The partition P_2 divides the interval $I_2 = [1, 6]$ into the two subintervals $[1, 3]$ and $[3, 6]$. Therefore, the partition $\mathbf{P} = (P_1, P_2)$ divides the rectangle $\mathbf{I}$ into the following six rectangles.

$$[-2, 0] \times [1, 3], \quad [0, 4] \times [1, 3], \quad [4, 9] \times [1, 3],$$
$$[-2, 0] \times [3, 6], \quad [0, 4] \times [3, 6], \quad [4, 9] \times [3, 6].$$

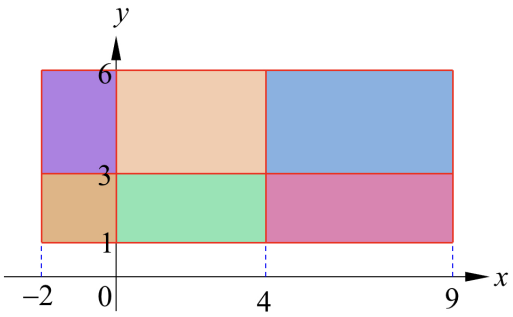


Figure 6.1: A partition of the rectangle $[-2, 9] \times [1, 6]$ given in Example 6.1.

Definition 6.2 Regular and Uniformly Regular Partitions

Let $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$ be a rectangle in $\mathbb{R}^n$. We say that $\mathbf{P} = (P_1, \ldots, P_n)$ is a *regular partition* of $\mathbf{I}$ if for each $1 \leq i \leq n$, P_i is a regular partition of $[a_i, b_i]$ into k_i intervals. We say that $\mathbf{P}$ is a *uniformly regular partition* of $\mathbf{I}$ into k^n rectangles if for each $1 \leq i \leq n$, P_i is a regular partition of $[a_i, b_i]$ into k intervals.

Example 6.2

Consider the rectangle $\mathbf{I} = [-2, 7] \times [-4, 8]$.

(a) The partition $\mathbf{P} = (P_1, P_2)$ where $P_1 = \{-2, 1, 4, 7\}$ and $P_2 = \{-4, -1, 2, 5, 8\}$ is a regular partition of $\mathbf{I}$.

(b) The partition $\mathbf{P} = (P_1, P_2)$ where $P_1 = \{-2, 1, 4, 7\}$ and $P_2 = \{-4, 0, 4, 8\}$ is a uniformly regular partition of $\mathbf{I}$ into $3^2 = 9$ rectangles.

The length of an interval $[a, b]$ is $b - a$. The area of a rectangle $[a, b] \times [c, d]$ is $(b - a) \times (d - c)$. In general, we define the volume of a closed rectangle of the form $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$ in $\mathbb{R}^n$ as follows.

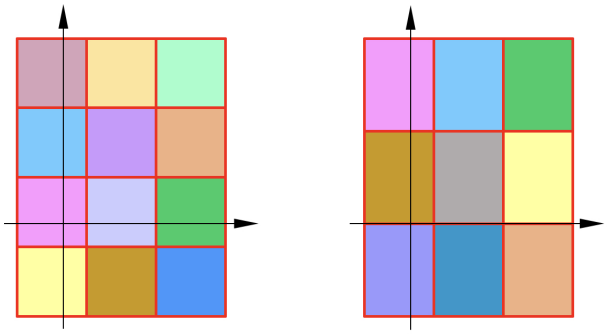


Figure 6.2: A regular and a uniformly regular partition of $[-2,7] \times [-4,8]$ discussed in Example 6.2.

Definition 6.3 Volume of a Rectangle

The volume of the closed rectangle $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$ is defined as the product of the lengths of all its edges. Namely,

$$\text{vol}\,(\mathbf{I}) = \prod_{i=1}^{n} (b_i - a_i).$$

Example 6.3

The volume of the rectangle $\mathbf{I} = [-2,9] \times [1,6]$ is

$$\text{vol}\,(\mathbf{I}) = 11 \times 5 = 55.$$

When $P = \{x_0, x_1, \ldots, x_k\}$ is a partition of $[a,b]$, it divides $[a,b]$ into k subintervals $J_1, \ldots, J_k$, where $J_i = [x_{i-1}, x_i]$. Notice that

$$\sum_{i=1}^{k} \text{vol}\,(J_i) = \sum_{i=1}^{k} (x_i - x_{i-1}) = b - a.$$

Assume that $\mathbf{P} = (P_1, \cdots, P_n)$ is a partition of the rectangle $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$ in $\mathbb{R}^n$. Then for $1 \leq i \leq n$, P_i is a partition of $[a_i, b_i]$. If P_i divides $[a_i, b_i]$ into the k_i

subintervals $J_{i,1}, J_{i,2}, \ldots, J_{i,k_i}$, then the collection of rectangles in the partition $\mathbf{P}$ is

$$\mathcal{J}_{\mathbf{P}} = \{J_{1,m_1} \times \cdots \times J_{n,m_n} \,|\, 1 \leq m_i \leq k_i \text{ for } 1 \leq i \leq n\}.$$

Notice that

$$\text{vol}\,(J_{1,m_1} \times \cdots \times J_{n,m_n}) = \text{vol}\,(J_{1,m_1}) \times \cdots \times \text{vol}\,(J_{n,m_n}).$$

From this, we obtain the sum of volumes formula:

$$\begin{aligned}
\sum_{\mathbf{J} \in \mathcal{J}_{\mathbf{P}}} \text{vol}\,(\mathbf{J}) &= \sum_{m_n=1}^{k_n} \cdots \sum_{m_1=1}^{k_1} \text{vol}\,(J_{1,m_1}) \times \cdots \times \text{vol}\,(J_{n,m_n}) \\
&= \left[\sum_{m_1=1}^{k_1} \text{vol}\,(J_{1,m_1})\right] \times \cdots \times \left[\sum_{m_n=1}^{k_n} \text{vol}\,(J_{n,m_n})\right] \\
&= (b_1 - a_1) \times \cdots \times (b_n - a_n) \\
&= \text{vol}\,(\mathbf{I}).
\end{aligned}$$

Proposition 6.1

Let $\mathbf{P}$ be a partition of $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$. Then the sum of the volumes of the rectangles $\mathbf{J}$ in the partition $\mathbf{P}$ is equal to the volume of the rectangle $\mathbf{I}$.

One of the motivations to define the integral $\int_{\mathbf{I}} f$ for a nonnegative function $f : \mathbf{I} \to \mathbb{R}$ is to find the volume bounded between the graph of f and the rectangle $\mathbf{I}$ in $\mathbb{R}^{n+1}$. To find the volume, we partition $\mathbf{I}$ into small rectangles, pick a point $\boldsymbol{\xi}_{\mathbf{J}}$ in each of these rectangles $\mathbf{J}$, and approximate the function on $\mathbf{J}$ as a constant given by the value $f(\boldsymbol{\xi}_{\mathbf{J}})$. The volume between the rectangle $\mathbf{J}$ and the graph of f over $\mathbf{J}$ is then approximated by $f(\boldsymbol{\xi}_{\mathbf{J}})\,\text{vol}\,(\mathbf{J})$. This leads us to the concept of Riemann sums.

If $\mathbf{P}$ is a partition of $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$, we say that A is a set of *intermediate points* for the partition $\mathbf{P}$ if $A = \{\boldsymbol{\xi}_{\mathbf{J}} \,|\, \mathbf{J} \in \mathcal{J}_{\mathbf{P}}\}$ is a subset of $\mathbf{I}$ indexed by $\mathcal{J}_{\mathbf{P}}$, such that $\boldsymbol{\xi}_{\mathbf{J}} \in \mathbf{J}$ for each $\mathbf{J} \in \mathcal{J}_{\mathbf{P}}$.

Definition 6.4 Riemann Sums

Let $\mathbf{I} = \prod_{i=1}^{n}[a_i, b_i]$, and let $f : \mathbf{I} \to \mathbb{R}$ be a function defined on $\mathbf{I}$. Given a partition $\mathbf{P}$ of $\mathbf{I}$, a set $A = \{\boldsymbol{\xi}_\mathbf{J} \,|\, \mathbf{J} \in \mathcal{J}_\mathbf{P}\}$ of intermediate points for the partition $\mathbf{P}$, the Riemann sum of f with respect to the partition $\mathbf{P}$ and the set of intermediate points $A = \{\boldsymbol{\xi}_\mathbf{J}\}$ is the sum

$$R(f, \mathbf{P}, A) = \sum_{\mathbf{J}\in\mathcal{J}_\mathbf{P}} f(\boldsymbol{\xi}_\mathbf{J})\text{vol}\,(\mathbf{J}).$$

Example 6.4

Let $\mathbf{I} = [-2, 9] \times [1, 6]$, and let $\mathbf{P} = (P_1, P_2)$ be the partition of $\mathbf{I}$ with $P_1 = \{-2, 0, 4, 9\}$ and $P_2 = \{1, 3, 6\}$. Let $f : \mathbf{I} \to \mathbb{R}$ be the function defined as $f(x, y) = x^2 + y$. Consider a set of intermediate points A as follows.

$\mathbf{J}$	$\boldsymbol{\xi}_\mathbf{J}$	$f(\boldsymbol{\xi}_\mathbf{J})$	vol $(\mathbf{J})$
$[-2, 0] \times [1, 3]$	$(-1, 1)$	2	4
$[-2, 0] \times [3, 6]$	$(0, 3)$	3	6
$[0, 4] \times [1, 3]$	$(1, 1)$	2	8
$[0, 4] \times [3, 6]$	$(2, 4)$	8	12
$[4, 9] \times [1, 3]$	$(4, 2)$	18	10
$[4, 9] \times [3, 6]$	$(9, 3)$	84	15

The Riemann sum $R(f, \mathbf{P}, A)$ is equal to

$$2 \times 4 + 3 \times 6 + 2 \times 8 + 8 \times 12 + 18 \times 10 + 84 \times 15 = 1578.$$

Example 6.5

If $f : \mathbf{I} \to \mathbb{R}$ is the constant function $f(\mathbf{x}) = c$, then for any partition $\mathbf{P}$ of $\mathbf{I}$ and any set of intermediate points $A = \{\boldsymbol{\xi}_\mathbf{J}\}$,

$$R(f, \mathbf{P}, A) = c\,\text{vol}\,(\mathbf{I}).$$

When $c > 0$, this is the volume of the rectangle $\mathbf{I} \times [0, c]$ in $\mathbb{R}^{n+1}$.

As in the single variable case, Darboux sums provide bounds for Riemann sums.

Definition 6.5 Darboux Sums

Let $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$, and let $f : \mathbf{I} \to \mathbb{R}$ be a *bounded* function defined on $\mathbf{I}$. Given a partition $\mathbf{P}$ of $\mathbf{I}$, let $\mathcal{J}_\mathbf{P}$ be the collection of rectangles in the partition $\mathbf{P}$. For each $\mathbf{J}$ in $\mathcal{J}_\mathbf{P}$, let

$$m_\mathbf{J} = \inf \{f(\mathbf{x}) \,|\, \mathbf{x} \in \mathbf{J}\} \quad \text{and} \quad M_\mathbf{J} = \sup \{f(\mathbf{x}) \,|\, \mathbf{x} \in \mathbf{J}\} .$$

The Darboux lower sum $L(f, \mathbf{P})$ and the Darboux upper sum $U(f, \mathbf{P})$ are defined as

$$L(f, \mathbf{P}) = \sum_{\mathbf{J} \in \mathcal{J}_\mathbf{P}} m_\mathbf{J} \text{ vol}\,(\mathbf{J}) \quad \text{and} \quad U(f, \mathbf{P}) = \sum_{\mathbf{J} \in \mathcal{J}_\mathbf{P}} M_\mathbf{J} \text{ vol}\,(\mathbf{J}).$$

Example 6.6

If $f : \mathbf{I} \to \mathbb{R}$ is the constant function $f(\mathbf{x}) = c$, then

$$L(f, \mathbf{P}) = c \,\text{vol}\,(\mathbf{I}) = U(f, \mathbf{P}) \qquad \text{for any partition } \mathbf{P} \text{ of } \mathbf{I}.$$

Example 6.7

Consider the function $f : \mathbf{I} \to \mathbb{R}$, $f(x, y) = x^2 + y$ defined in Example 6.4, where $\mathbf{I} = [-2, 9] \times [1, 6]$. For the partition $\mathbf{P} = (P_1, P_2)$ with $P_1 = \{-2, 0, 4, 9\}$ and $P_2 = \{1, 3, 6\}$, we have the followings.

$\mathbf{J}$	$m_\mathbf{J}$	$M_\mathbf{J}$	vol $(\mathbf{J})$
$[-2, 0] \times [1, 3]$	$0^2 + 1 = 1$	$(-2)^2 + 3 = 7$	4
$[-2, 0] \times [3, 6]$	$0^2 + 3 = 3$	$(-2)^2 + 6 = 10$	6
$[0, 4] \times [1, 3]$	$0^2 + 1 = 1$	$4^2 + 3 = 19$	8
$[0, 4] \times [3, 6]$	$0^2 + 3 = 3$	$4^2 + 6 = 22$	12
$[4, 9] \times [1, 3]$	$4^2 + 1 = 17$	$9^2 + 3 = 84$	10
$[4, 9] \times [3, 6]$	$4^2 + 3 = 19$	$9^2 + 6 = 87$	15

Therefore, the Darboux lower sum is

$$L(f, \mathbf{P}) = 1 \times 4 + 3 \times 6 + 1 \times 8 + 3 \times 12 + 17 \times 10 + 19 \times 15 = 521;$$

while the Darboux upper sum is

$$U(f, \mathbf{P}) = 7 \times 4 + 10 \times 6 + 19 \times 8 + 22 \times 12 + 84 \times 10 + 87 \times 15 = 2649.$$

Notice that we can only define Darboux sums if the function $f : \mathbf{I} \to \mathbb{R}$ is bounded. This means that there are constants m and M such that

$$m \leq f(x) \leq M \qquad \text{for all } \mathbf{x} \in \mathbf{I}.$$

If $\mathbf{P}$ is a partition of the rectangle $\mathbf{I}$, and $\mathbf{J}$ is a rectangle in the partition $\mathbf{P}$, $\boldsymbol{\xi}_J$ is a point in $\mathbf{J}$, then

$$m \leq m_{\mathbf{J}} \leq f(\boldsymbol{\xi}_{\mathbf{J}}) \leq M_{\mathbf{J}} \leq M.$$

Multipluying throughout by $\text{vol}\,(\mathbf{J})$ and summing over $\mathbf{J} \in \mathcal{J}_{\mathbf{P}}$, we obtain the following.

Proposition 6.2

Let $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$, and let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function defined on $\mathbf{I}$. If

$$m \leq f(\mathbf{x}) \leq M \qquad \text{for all } \mathbf{x} \in \mathbf{I},$$

then for any partition $\mathbf{P}$ of $\mathbf{I}$, and for any choice of intermediate points $A = \{\boldsymbol{\xi}_{\mathbf{J}}\}$ for the partition $\mathbf{P}$, we have

$$m\,\text{vol}\,(\mathbf{I}) \;\leq\; L(f, \mathbf{P}) \;\leq\; R(f, \mathbf{P}, A) \;\leq\; U(f, \mathbf{P}) \;\leq\; M\,\text{vol}\,(\mathbf{I}).$$

To study the behaviour of the Darboux sums when we modify the partitions, we first extend the concept of refinement of a partition to rectangles in $\mathbb{R}^n$. Recall that if P and P^* are partitions of the interval $[a, b]$, P^* is a refinement of P if each partition point of P is also a partition point of P^*.

Definition 6.6 Refinement of a Partition

Let $\mathbf{I} = \prod_{i=1}^{n}[a_i, b_i]$, and let $\mathbf{P} = (P_1, \ldots, P_n)$ and $\mathbf{P}^* = (P_1^*, \ldots, P_n^*)$ be partitions of $\mathbf{I}$. We say that $\mathbf{P}^*$ is a refinement of $\mathbf{P}$ if for each $1 \leq i \leq n$, P_i^* is a refinement of P_i.

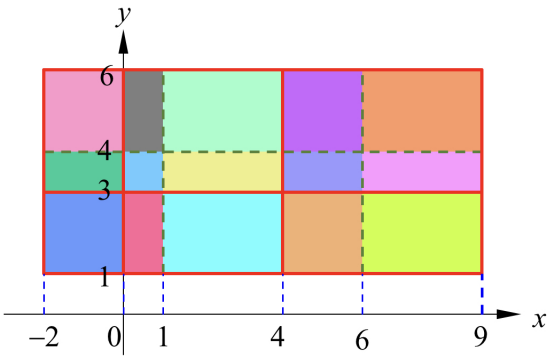


Figure 6.3: A refinement of the partition of the rectangle $[-2, 9] \times [1, 6]$ given in Figure 6.1.

Example 6.8

Let us consider the partition $\mathbf{P} = (P_1, P_2)$ of the rectangle $\mathbf{I} = [-2, 9] \times [1, 6]$ given in Example 6.1, with $P_1 = \{-2, 0, 4, 9\}$ and $P_2 = \{1, 3, 6\}$. Let $P_1^* = \{-2, 0, 1, 4, 6, 9\}$ and $P_2^* = \{1, 3, 4, 6\}$. Then $\mathbf{P}^* = (P_1^*, P_2^*)$ is a refinement of $\mathbf{P}$.

If the partition $\mathbf{P}^*$ is a refinement of the partition $\mathbf{P}$, then for each $\mathbf{J}$ in $\mathcal{J}_\mathbf{P}$, $\mathbf{P}^*$ induces a partition of $\mathbf{J}$, which we denote by $\mathbf{P}^*(\mathbf{J})$.

Example 6.9

The partition $\mathbf{P}^*$ in Example 6.8 induces the partition $\mathbf{P}^*(\mathbf{J}) = (P_1^*(\mathbf{J}), P_2^*(\mathbf{J}))$ of the rectangle $\mathbf{J} = [0, 4] \times [3, 6]$, where $P_1^*(\mathbf{J}) = \{0, 1, 4\}$ and $P_2^*(\mathbf{J}) = \{3, 4, 6\}$. The partition $\mathbf{P}^*(\mathbf{J})$ divides the rectangle $\mathbf{J}$ into 4 rectangles, as shown in Figure 6.3.

If the partition $\mathbf{P}^*$ is a refinement of the partition $\mathbf{P}$, then the collection of rectangles in $\mathbf{P}^*$ is the union of the collection of rectangles in $\mathbf{P}^*(\mathbf{J})$ when $\mathbf{J}$ ranges over the collection of rectangles in $\mathbf{P}$. Namely,

$$\mathcal{J}_{\mathbf{P}^*} = \bigcup_{\mathbf{J}\in\mathcal{J}_{\mathbf{P}}} \mathcal{J}_{\mathbf{P}^*(\mathbf{J})}.$$

Using this, we can deduce the following.

Proposition 6.3

Let $\mathbf{I} = \prod_{i=1}^{n}[a_i, b_i]$, and let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function defined on $\mathbf{I}$. If $\mathbf{P}$ and $\mathbf{P}^*$ are partitions of $\mathbf{I}$ and $\mathbf{P}^*$ is a refinement of $\mathbf{P}$, then

$$L(f, \mathbf{P}^*) = \sum_{\mathbf{J}\in\mathcal{J}_{\mathbf{P}}} L(f, \mathbf{P}^*(\mathbf{J})), \qquad U(f, \mathbf{P}^*) = \sum_{\mathbf{J}\in\mathcal{J}_{\mathbf{P}}} U(f, \mathbf{P}^*(\mathbf{J})).$$

From this, we can show that a refinement improves the Darboux sums, in the sense that a lower sum increases, and an upper sum decreases.

Theorem 6.4

Let $\mathbf{I} = \prod_{i=1}^{n}[a_i, b_i]$, and let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function defined on $\mathbf{I}$. If $\mathbf{P}$ and $\mathbf{P}^*$ are partitions of $\mathbf{I}$ and $\mathbf{P}^*$ is a refinement of $\mathbf{P}$, then

$$L(f, \mathbf{P}) \leq L(f, \mathbf{P}^*) \leq U(f, \mathbf{P}^*) \leq U(f, \mathbf{P}).$$

Proof

For each rectangle $\mathbf{J}$ in the partition $\mathbf{P}$,

$$m_{\mathbf{J}} \leq f(\mathbf{x}) \leq M_{\mathbf{J}} \qquad \text{for all } \mathbf{x} \in \mathbf{J}.$$

Applying Proposition 6.2 to the function $f : \mathbf{J} \to \mathbb{R}$ and the partition $\mathbf{P}^*(\mathbf{J})$, we find that

$$m_{\mathbf{J}} \text{ vol } (\mathbf{J}) \leq L(f, \mathbf{P}^*(\mathbf{J})) \leq U(f, \mathbf{P}^*(\mathbf{J})) \leq M_{\mathbf{J}} \text{ vol } (\mathbf{J}).$$

Summing over $\mathbf{J} \in \mathcal{J}_{\mathbf{P}}$, we find that

$$L(f, \mathbf{P}) \leq \sum_{\mathbf{J} \in \mathcal{J}_{\mathbf{P}}} L(f, \mathbf{P}^*(\mathbf{J})) \leq \sum_{\mathbf{J} \in \mathcal{J}_{\mathbf{P}}} U(f, \mathbf{P}^*(\mathbf{J})) \leq U(f, \mathbf{P}).$$

The assertion follows from Proposition 6.3.

It is difficult to visualize the Darboux sums with a multivariable functions. Hence, we illustrate refinements improve Darboux sums using single variable functions, as shown in Figure 6.4 and Figure 6.5.

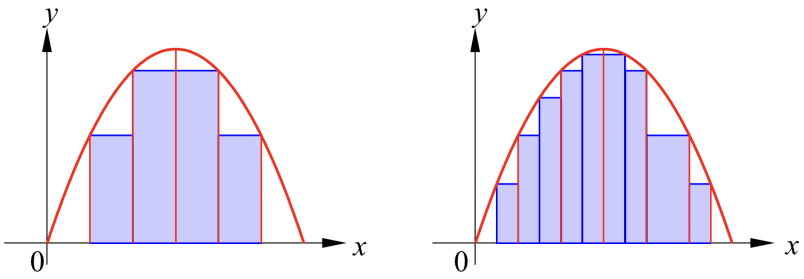


Figure 6.4: A refinement of the partition increases the Darboux lower sum.

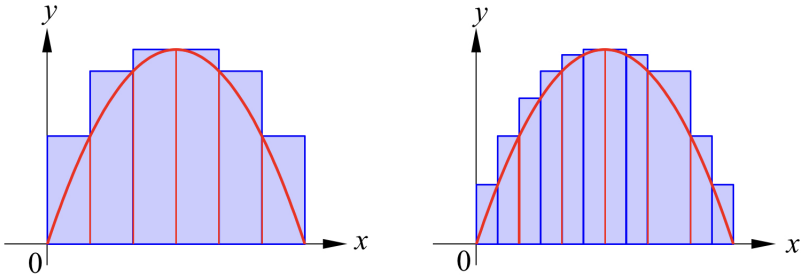


Figure 6.5: A refinement of the partition decreases the Darboux upper sum.

As a consequence of Theorem 6.4, we can prove the following.

Corollary 6.5

Let $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$, and let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function defined on $\mathbf{I}$. For any two partitions $\mathbf{P}_1$ and $\mathbf{P}_2$ of $\mathbf{I}$,

$$L(f, \mathbf{P}_1) \leq U(f, \mathbf{P}_2).$$

Proof

Let $\mathbf{P}_1 = (P_{1,1}, P_{1,2}, \ldots, P_{1,n})$ and $\mathbf{P}_2 = (P_{2,1}, P_{2,2}, \ldots, P_{2,n})$. For $1 \leq i \leq n$, let P_i^* be the common refinement of $P_{1,i}$ and $P_{2,i}$ obtained by taking the union of the partition points in $P_{1,i}$ and $P_{2,i}$. Then $\mathbf{P}^* = (P_1^*, \ldots, P_n^*)$ is a common refinement of the partitions $\mathbf{P}_1$ and $\mathbf{P}_2$. By Theorem 6.4,

$$L(f, \mathbf{P}_1) \leq L(f, \mathbf{P}^*) \leq U(f, \mathbf{P}^*) \leq U(f, \mathbf{P}_2).$$

Now we define lower and upper integrals of a bounded function $f : \mathbf{I} \to \mathbb{R}$.

Definition 6.7 Lower Integrals and Upper Integrals

Let $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$, and let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function defined on $\mathbf{I}$. Let $S_L(f)$ be the set of Darboux lower sums of f, and let $S_U(f)$ be the set of Darboux upper sums of f.

1. The lower integral of f, denoted by $\underline{\int_{\mathbf{I}}} f$, is defined as the least upper bound of the Darboux lower sums.

$$\underline{\int_{\mathbf{I}}} f = \sup S_L(f) = \sup \{L(f, \mathbf{P}) \mid \mathbf{P} \text{ is a partition of } \mathbf{I}\} .$$

2. The upper integral of f, denoted by $\overline{\int_{\mathbf{I}}} f$, is defined as the greatest lower bound of the Darboux upper sums.

$$\overline{\int_{\mathbf{I}}} f = \inf S_U(f) = \inf \{U(f, \mathbf{P}) \mid \mathbf{P} \text{ is a partition of } \mathbf{I}\} .$$

Example 6.10

If $f : \mathbf{I} \to \mathbb{R}$ is the constant function $f(\mathbf{x}) = c$, then for any partition $\mathbf{P}$ of $\mathbf{I}$,

$$L(f, \mathbf{P}) = c \operatorname{vol}(\mathbf{I}) = U(f, \mathbf{P}).$$

Therefore, both $S_L(f)$ and $S_U(f)$ are the one-element set $\{c \operatorname{vol}(\mathbf{I})\}$. This shows that

$$\underline{\int_{\mathbf{I}}} f = \overline{\int_{\mathbf{I}}} f = c \operatorname{vol}(\mathbf{I}).$$

For a constant function, the lower integral and the upper integral are the same. For a general bounded funtion, we have the following.

Theorem 6.6

Let $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$, and let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function defined on $\mathbf{I}$. Then we have

$$\underline{\int_{\mathbf{I}}} f \leq \overline{\int_{\mathbf{I}}} f.$$

Proof

By Corollary 6.5, every element of $S_L(f)$ is less than or equal to any element of $S_U(f)$. This implies that

$$\underline{\int_{\mathbf{I}}} f = \sup S_L(f) \leq \inf S_U(f) = \overline{\int_{\mathbf{I}}} f.$$

Example 6.11 The Dirichlet's Function

Let $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$, and let $f : \mathbf{I} \to \mathbb{R}$ be the function defined as

$$f(\mathbf{x}) = \begin{cases} 1, & \text{if all components of } \mathbf{x} \text{ are rational,} \\ 0, & \text{otherwise.} \end{cases}$$

This is known as the Dirichlet's function. Find the lower inegral and the upper integral of $f : \mathbf{I} \to \mathbb{R}$.

Solution

Let $\mathbf{P} = (P_1, \ldots, P_n)$ be a partition of $\mathbf{I}$. A rectangle $\mathbf{J}$ in the partition $\mathbf{P}$ can be written in the form $\mathbf{J} = \prod_{i=1}^{n} [u_i, v_i]$. By denseness of rational numbers and irrational numbers, there exist a rational number α_i and an irrational number β_i in (u_i, v_i). Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)$. Then $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are points in $\mathbf{J}$, and

$$0 = f(\boldsymbol{\beta}) \leq f(\mathbf{x}) \leq f(\boldsymbol{\alpha}) = 1 \qquad \text{for all } \mathbf{x} \in \mathbf{J}.$$

Therefore,

$$m_{\mathbf{J}} = \inf_{\mathbf{x} \in \mathbf{J}} f(\mathbf{x}) = 0, \qquad M_{\mathbf{J}} = \sup_{\mathbf{x} \in \mathbf{J}} f(\mathbf{x}) = 1.$$

It follows that

$$L(f, \mathbf{P}) = \sum_{\mathbf{J} \in \mathcal{J}_{\mathbf{P}}} m_{\mathbf{J}} \operatorname{vol}(\mathbf{J}) = 0,$$

$$U(f, \mathbf{P}) = \sum_{\mathbf{J} \in \mathcal{J}_{\mathbf{P}}} M_{\mathbf{J}} \operatorname{vol}(\mathbf{J}) = \sum_{\mathbf{J} \in \mathcal{J}_{\mathbf{P}}} \operatorname{vol}(\mathbf{J}) = \operatorname{vol}(\mathbf{I}).$$

Therefore,

$$S_L(f) = \{0\}, \quad \text{while} \quad S_U(f) = \{\operatorname{vol}(\mathbf{I})\}.$$

This shows that the lower inegral and the upper integral of $f : \mathbf{I} \to \mathbb{R}$ are given respectively by

$$\underline{\int_{\mathbf{I}}} f = 0 \quad \text{and} \quad \overline{\int_{\mathbf{I}}} f = \operatorname{vol}(\mathbf{I}).$$

As we mentioned before, one of the motivations to define the integral $f : \mathbf{I} \to \mathbb{R}$ is to calculate volumes. Given that $f : \mathbf{I} \to \mathbb{R}$ is a nonnegative continuous function defined on the rectangle $\mathbf{I}$ in $\mathbb{R}^n$, let

$$S = \{(\mathbf{x}, y) \,|\, \mathbf{x} \in \mathbf{I}, 0 \leq y \leq f(\mathbf{x})\},$$

which is the solid bounded between $\mathbf{I}$ and the graph of f. It is reasonable to expect that S has a volume, which we denote by $\operatorname{vol}(S)$. We want to define the integral

$\int_{\mathbf{I}} f$ so that it gives vol (S). Notice that if $\mathbf{P}$ is a partition of $\mathbf{I}$, then the Darboux lower sum

$$L(f, \mathbf{P}) = \sum_{\mathbf{J} \in \mathcal{J}_{\mathbf{P}}} m_{\mathbf{J}} \text{ vol } (\mathbf{J})$$

is the sum of volumes of the collection of rectangles

$$\{\mathbf{J} \times [0, m_{\mathbf{J}}] \,|\, \mathbf{J} \in \mathcal{J}_{\mathbf{P}}\}$$

in $\mathbb{R}^{n+1}$, each of which is contained in S. Since any two of these rectangles can only intersect on the boundaries, it is reasonable to expect that

$$L(f, \mathbf{P}) \leq \text{vol } (S).$$

Similarly, the Darboux upper sum

$$U(f, \mathbf{P}) = \sum_{\mathbf{J} \in \mathcal{J}_{\mathbf{P}}} M_{\mathbf{J}} \text{ vol } (\mathbf{J})$$

is the sum of volumes of the collection of rectangles

$$\{\mathbf{J} \times [0, M_{\mathbf{J}}] \,|\, \mathbf{J} \in \mathcal{J}_{\mathbf{P}}\}$$

in $\mathbb{R}^{n+1}$, the union of which contains S. Therefore, it is reasonable to expect that

$$\text{vol } (S) \leq U(f, \mathbf{P}).$$

Hence, the volume of S should be a number between $L(f, \mathbf{P})$ and $U(f, \mathbf{P})$ for any partition $\mathbf{P}$. To make the volume well-defined, there should be only one number between $L(f, \mathbf{P})$ and $U(f, \mathbf{P})$ for all partitions $\mathbf{P}$. By definition, any number between the lower integral and the upper integral is in between $L(f, \mathbf{P})$ and $U(f, \mathbf{P})$ for any partition $\mathbf{P}$. Hence, to have the volume well-defined, we must require the lower integral and the upper integral to be the same. This motivates the following definition of integrability for a general bounded function.

Definition 6.8 Riemann integrability

Let $\mathbf{I} = \prod_{i=1}^{n}[a_i, b_i]$, and let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function defined on $\mathbf{I}$. We say that $f : \mathbf{I} \to \mathbb{R}$ is Riemann integrable, or simply integrable, if

$$\underline{\int_{\mathbf{I}}} f = \overline{\int_{\mathbf{I}}} f.$$

In this case, we define the integral of f over the rectangle $\mathbf{I}$ as

$$\int_{\mathbf{I}} f = \underline{\int_{\mathbf{I}}} f = \overline{\int_{\mathbf{I}}} f.$$

It is the unique number larger than or equal to all Darboux lower sums, and smaller than or equal to all Darboux upper sums.

Example 6.12

Example 6.10 says that a constant function $f : \mathbf{I} \to \mathbb{R}$, $f(\mathbf{x}) = c$ is integrable and

$$\int_{\mathbf{I}} f = c \operatorname{vol}(\mathbf{I}).$$

Example 6.13

The Dirichlet's function defined in Example 6.11 is not Riemann integrable since the lower integral and the upper integral are not equal.

Leibniz Notation for Riemann Integrals

The Leibniz notation of the Riemann integral of $f : \mathbf{I} \to \mathbb{R}$ is

$$\int_{\mathbf{I}} f(\mathbf{x}) d\mathbf{x}, \quad \text{or equivalently,} \quad \int_{\mathbf{I}} f(x_1, \ldots, x_n) dx_1 \cdots dx_n.$$

As in the single variable case, there are some criteria for Riemann integrability which follows directly from the criteria that the lower integral and the upper integral are the same.

Theorem 6.7

Let $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$, and let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function defined on $\mathbf{I}$. The following are equivalent.

(a) The function $f : \mathbf{I} \to \mathbb{R}$ is Riemann integrable.

(b) For every $\varepsilon > 0$, there is a partition $\mathbf{P}$ of the rectangle $\mathbf{I}$ such that

$$U(f, \mathbf{P}) - L(f, \mathbf{P}) < \varepsilon.$$

We define an Archimedes sequence of partitions exactly the same as in the single variable case.

Definition 6.9 Archimedes Sequence of Partitions

Let $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$, and let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function defined on $\mathbf{I}$. If $\{\mathbf{P}_k\}$ is a sequence of partitions of the rectangle $\mathbf{I}$ such that

$$\lim_{k \to \infty} (U(f, \mathbf{P}_k) - L(f, \mathbf{P}_k)) = 0,$$

we call $\{\mathbf{P}_k\}$ an Archimedes sequence of partitions for the function f.

Then we have the following theorem.

Theorem 6.8 The Archimedes-Riemann Theorem

Let $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$, and let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function defined on $\mathbf{I}$. The function $f : \mathbf{I} \to \mathbb{R}$ is Riemann integrable if and only if f has an Archimedes sequence of partitions $\{\mathbf{P}_k\}$. In this case, the integral $\int_{\mathbf{I}} f$ can be computed by

$$\int_{\mathbf{I}} f = \lim_{k \to \infty} L(f, \mathbf{P}_k) = \lim_{k \to \infty} U(f, \mathbf{P}_k).$$

A candidate for an Archimedes sequence of partitions is the sequence $\{\mathbf{P}_k\}$,

where $\mathbf{P}_k$ is the uniformly regular partition of $\mathbf{I}$ into k^n rectangles.

Example 6.14

Let $\mathbf{I} = [0,1] \times [0,1]$. Consider the function $f : \mathbf{I} \to \mathbb{R}$ defined as

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

For $k \in \mathbb{Z}^+$, let $\mathbf{P}_k$ be the uniformly regular partition of $\mathbf{I}$ into k^2 rectangles.

(a) For each $k \in \mathbb{Z}^+$, compute the Darboux lower sum $L(f, \mathbf{P}_k)$ and the Darboux upper sum $U(f, \mathbf{P}_k)$.

(b) Show that $f : \mathbf{I} \to \mathbb{R}$ is Riemann integrable and find the integral $\displaystyle\int_{\mathbf{I}} f$.

Solution

Fixed $k \in \mathbb{Z}^+$, let $P_k = \{u_0, u_1, \ldots, u_k\}$, where $u_i = \dfrac{i}{k}$ for $0 \leq i \leq k$. Then $\mathbf{P}_k = (P_k, P_k)$, and it divides $\mathbf{I} = [0,1] \times [0,1]$ into the k^2 rectangles $\mathbf{J}_{i,j}$, $1 \leq i \leq k$, $1 \leq j \leq k$, where $\mathbf{J}_{i,j} = [u_{i-1}, u_i] \times [u_{j-1}, u_j]$. We have

$$\text{vol}\,(\mathbf{J}_{i,j}) = \frac{1}{k^2}.$$

Let

$$m_{i,j} = \inf_{(x,y) \in \mathbf{J}_{i,j}} f(x,y) \quad \text{and} \quad M_{i,j} = \sup_{(x,y) \in \mathbf{J}_{i,j}} f(x,y).$$

Notice that if $i < j - 1$, then

$$x \leq u_i < u_{j-1} \leq y \qquad \text{for all } (x,y) \in \mathbf{J}_{i,j}.$$

Hence,

$$f(x,y) = 0 \qquad \text{for all } (x,y) \in \mathbf{J}_{i,j}.$$

This implies that

$$m_{i,j} = M_{i,j} = 0 \qquad \text{when } i < j - 1.$$

If $i \geq j+1$, then

$$x \geq u_{i-1} \geq u_j \geq y \qquad \text{for all } (x,y) \in \mathbf{J}_{i,j}.$$

Hence,

$$f(x,y) = 1 \qquad \text{for all } (x,y) \in \mathbf{J}_{i,j}.$$

This implies that

$$m_{i,j} = M_{i,j} = 1 \qquad \text{when } i \geq j+1.$$

When $i = j-1$, if (x,y) is in $\mathbf{J}_{i,j}$,

$$x \leq u_i = u_{j-1} \leq y,$$

and $x = y$ if and only if (x,y) is the point (u_i, u_{j-1}). Hence, $f(x,y) = 0$ for all $(x,y) \in \mathbf{J}_{i,j}$, except for $(x,y) = (u_i, u_{j-1})$, where $f(u_i, u_{j-1}) = 1$. Hence,

$$m_{i,j} = 0, \quad M_{i,j} = 1 \qquad \text{when } i = j-1.$$

When $i = j$, $0 \leq f(x,y) \leq 1$ for all $(x,y) \in \mathbf{J}_{i,j}$. Since (u_{i-1}, u_j) and (u_i, u_j) are in $\mathbf{J}_{i,j}$, and $f(u_{i-1}, u_j) = 0$ while $f(u_i, u_j) = 1$, we find that

$$m_{i,j} = 0, \quad M_{i,j} = 1 \qquad \text{when } i = j.$$

It follows that

$$\begin{aligned} L(f, \mathbf{P}_k) &= \sum_{i=1}^{k} \sum_{j=1}^{k} m_{i,j} \operatorname{vol}(\mathbf{J}_{i,j}) = \sum_{i=2}^{k} \sum_{j=1}^{i-1} \frac{1}{k^2} \\ &= \frac{1}{k^2} \sum_{i=2}^{k} (i-1) = \frac{1}{k^2} \sum_{i=1}^{k-1} i = \frac{k(k-1)}{2k^2}. \\ U(f, \mathbf{P}_k) &= \sum_{i=1}^{k} \sum_{j=1}^{k} M_{i,j} \operatorname{vol}(\mathbf{J}_{i,j}) = \sum_{i=1}^{k-1} \sum_{j=1}^{i+1} \frac{1}{k^2} + \sum_{j=1}^{k} \frac{1}{k^2} \\ &= \frac{1}{k} + \frac{1}{k^2} \sum_{i=1}^{k-1} (i+1) = \frac{1}{k^2} \left(\frac{k(k+1)}{2} - 1 + k \right) = \frac{k^2 + 3k - 2}{2k^2}. \end{aligned}$$

Since

$$U(f, \mathbf{P}_k) - L(f, \mathbf{P}_k) = \frac{2k-1}{k^2} \qquad \text{for all } k \in \mathbb{Z}^+,$$

we find that

$$\lim_{k\to\infty} \left(U(f, \mathbf{P}_k) - L(f, \mathbf{P}_k)\right) = 0.$$

Hence, $\{\mathbf{P}_k\}$ is an Archimedes sequence of partitions for f. By the Arichimedes-Riemann theorem, $f : \mathbf{I} \to \mathbb{R}$ is Riemann integrable, and

$$\int_{\mathbf{I}} f = \lim_{k\to\infty} L(f, \mathbf{P}_k) = \lim_{k\to\infty} \frac{k(k-1)}{2k^2} = \frac{1}{2}.$$

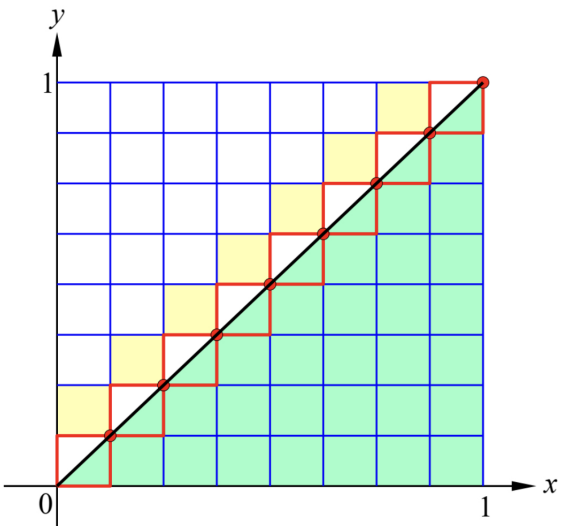


Figure 6.6: This figure illustrates the different cases considered in Example 6.14 when $k = 8$.

As in the single variable case, there is an equivalent definition for Riemann integrability using Riemann sums.

For a partition $P = \{x_0, x_1, \ldots, x_k\}$ of an interval $[a, b]$, we define the gap of the partition P as

$$|P| = \max\left\{x_i - x_{i-1} \,|\, 1 \le i \le k\right\}.$$

For a closed rectangle $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$, we replace the length $x_i - x_{i-1}$ of an interval in the partition by the diameter of a rectangle in the partition. Recall that the

diameter of a rectangle $\mathbf{J} = \prod_{i=1}^{n}[u_i, v_i]$ is

$$\operatorname{diam} \mathbf{J} = \sqrt{(v_1 - u_1)^2 + \cdots + (v_n - u_n)^2}.$$

Definition 6.10 Gap of a Partition

Let $\mathbf{P}$ be a partition of the rectangle $\mathbf{I} = \prod_{i=1}^{n}[a_i, b_i]$. Then the gap of the partition $\mathbf{P}$ is defined as

$$|\mathbf{P}| = \max\{\operatorname{diam} \mathbf{J} \mid \mathbf{J} \in \mathcal{J}_{\mathbf{P}}\}.$$

Example 6.15

Find the gap of the partition $\mathbf{P} = (P_1, P_2)$ of the rectangle $\mathbf{I} = [-2, 9] \times [1, 6]$ defined in Example 6.1, where $P_1 = \{-2, 0, 4, 9\}$ and $P_2 = \{1, 3, 6\}$.

Solution

The length of the three invervals in the partition $P_1 = \{-2, 0, 4, 9\}$ of the interval $[-2, 9]$ are $2, 4$ and 5 respectively. The lengths of the two intervals in the partition $P_2 = \{1, 3, 6\}$ of the interval $[1, 6]$ are 2 and 3 respectively. Therefore, the diameters of the 6 rectangles in the partition $\mathbf{P}$ are

$$\sqrt{2^2 + 2^2}, \quad \sqrt{4^2 + 2^2}, \quad \sqrt{5^2 + 2^2},$$
$$\sqrt{2^2 + 3^2}, \quad \sqrt{4^2 + 3^2}, \quad \sqrt{5^2 + 3^2}.$$

From this, we see that the gap of $\mathbf{P}$ is $\sqrt{5^2 + 3^2} = \sqrt{34}$.

In the example above, notice that $|P_1| = 5$ and $|P_2| = 3$. In general, it is not difficult to see the following.

Proposition 6.9

Let $\mathbf{P} = (P_1, \ldots, P_n)$ be a partition of the closed rectangle $\mathbf{I} = \prod_{i=1}^{n}[a_i, b_i]$. Then

$$|\mathbf{P}| = \sqrt{|P_1|^2 + \cdots + |P_n|^2}.$$

The following theorem gives equivalent definitions of Riemann integrability of a bounded function.

Theorem 6.10 Equivalent Definitions for Riemann Integrability

Let $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$, and let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function defined on $\mathbf{I}$. The following three statements are equivalent for saying that $f : \mathbf{I} \to \mathbb{R}$ is Riemann integrable.

(a) The lower integral and the upper integral are the same. Namely,

$$\underline{\int_{\mathbf{I}}} f = \overline{\int_{\mathbf{I}}} f.$$

(b) There exists a number I that satisfies the following. For any $\varepsilon > 0$, there exists a $\delta > 0$ such that if $\mathbf{P}$ is a partition of the rectangle $\mathbf{I}$ with $|\mathbf{P}| < \delta$, then

$$|R(f, \mathbf{P}, A) - I| < \varepsilon$$

for any choice of intermediate points $A = \{\boldsymbol{\xi}_\mathbf{J}\}$ for the partition $\mathbf{P}$.

(c) For any $\varepsilon > 0$, there exists a $\delta > 0$ such that if $\mathbf{P}$ is a partition of the rectangle $\mathbf{I}$ with $|\mathbf{P}| < \delta$, then

$$U(f, \mathbf{P}) - L(f, \mathbf{P}) < \varepsilon.$$

The most useful definition is in fact the second one in terms of Riemann sums. It says that a bounded function $f : \mathbf{I} \to \mathbb{R}$ is Riemann integrable if the limit

$$\lim_{|\mathbf{P}| \to 0} R(f, \mathbf{P}, A)$$

exists. As a consequence of Theorem 6.10, we have the following.

Theorem 6.11

Let $\mathbf{I} = \prod_{i=1}^{n}[a_i, b_i]$, and let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function defined on $\mathbf{I}$. If $f : \mathbf{I} \to \mathbb{R}$ is Riemann integrable, then for any sequence $\{\mathbf{P}_k\}$ of partitions of $\mathbf{I}$ satisfying

$$\lim_{k\to\infty} |\mathbf{P}_k| = 0,$$

we have

(i) $\displaystyle\int_{\mathbf{I}} f = \lim_{k\to\infty} L(f, \mathbf{P}_k) = \lim_{k\to\infty} U(f, \mathbf{P}_k).$

(ii) $\displaystyle\int_{\mathbf{I}} f = \lim_{k\to\infty} R(f, \mathbf{P}_k, A_k)$, where for each $k \in \mathbb{Z}^+$, A_k is a choice of intermediate points for the partition $\mathbf{P}_k$.

The proof is exactly the same as the single variable case. The contrapositive of Theorem 6.11 gives the following.

Theorem 6.12

Let $\mathbf{I} = \prod_{i=1}^{n}[a_i, b_i]$, and let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function defined on $\mathbf{I}$. Assume that $\{\mathbf{P}_k\}$ is a sequence of partitions of $\mathbf{I}$ such that

$$\lim_{k\to\infty} |\mathbf{P}_k| = 0.$$

(a) If for each $k \in \mathbb{Z}^+$, there exists a choice of intermediate points A_k for the partition $\mathbf{P}_k$ such that the limit $\lim_{k\to\infty} R(f, \mathbf{P}_k, A_k)$ does not exist, then $f : \mathbf{I} \to \mathbb{R}$ is not Riemann integrable.

(b) If for each $k \in \mathbb{Z}^+$, there exist two choices of intermediate points A_k and B_k for the partition $\mathbf{P}_k$ so that the two limits $\lim_{k\to\infty} R(f, \mathbf{P}_k, A_k)$ and $\lim_{k\to\infty} R(f, \mathbf{P}_k, B_k)$ are not the same, then $f : \mathbf{I} \to \mathbb{R}$ is not Riemann integrable.

Theorem 6.12 is useful for justifying that a bounded function is not Riemann integrable, without having to compute the lower integral or the upper integral. To

apply this theorem, we usually consider the sequence of partitions $\{\mathbf{P}_k\}$, where $\mathbf{P}_k$ is the uniformly regular partition of $\mathbf{I}$ into k^n rectangles.

Example 6.16

Let $\mathbf{I} = [0,1] \times [0,1]$, and let $f : \mathbf{I} \to \mathbb{R}$ be the function defined as

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

Show that $f : \mathbf{I} \to \mathbb{R}$ is not Riemann integrable.

Solution

For $k \in \mathbb{Z}^+$, let $\mathbf{P}_k$ be the uniformly regular partition of $\mathbf{I}$ into k^2 rectangles. Then $\mathbf{P}_k = (P_k, P_k)$, where $P_k = \{u_0, u_1, \ldots, u_k\}$ with $u_i = \dfrac{i}{k}$ when $0 \le i \le k$. Notice that $|\mathbf{P}_k| = \dfrac{\sqrt{2}}{k}$, and so $\lim\limits_{k\to\infty} \mathbf{P}_k = 0$.
The partition $\mathbf{P}_k$ divides the square $\mathbf{I}$ into k^2 squares $\mathbf{J}_{i,j}$, $1 \le i \le k$, $1 \le j \le k$, where $\mathbf{J}_{i,j} = [u_{i-1}, u_i] \times [u_{j-1}, u_j]$. For $1 \le i \le k$, since irrational numbers are dense, there is an irrational number c_i in the interval (u_{i-1}, u_i). For $1 \le i \le k$, $1 \le j \le k$, let $\boldsymbol{\alpha}_{i,j}$ and $\boldsymbol{\beta}_{i,j}$ be the points in $\mathbf{J}_{i,j}$ given respectively by

$$\boldsymbol{\alpha}_{i,j} = (u_i, u_j), \qquad \boldsymbol{\beta}_{i,j} = (c_i, u_j).$$

Then

$$f(\boldsymbol{\alpha}_{i,j}) = 0, \qquad f(\boldsymbol{\beta}_{i,j}) = u_j.$$

Let $A_k = \{\boldsymbol{\alpha}_{i,j}\}$ and $B_k = \{\boldsymbol{\beta}_{i,j}\}$. Then the Riemann sums $R(f, \mathbf{P}_k, A_k)$ and $R(f, \mathbf{P}_k, B_k)$ are given respectively by

$$R(f, \mathbf{P}_k, A_k) = \sum_{i=1}^{k} \sum_{j=1}^{k} f(\boldsymbol{\alpha}_{i,j}) \operatorname{vol}(\mathbf{J}_{i,j}) = 0,$$

and

$$R(f, \mathbf{P}_k, B_k) = \sum_{i=1}^{k} \sum_{j=1}^{k} f(\boldsymbol{\beta}_{i,j}) \text{ vol } (\mathbf{J}_{i,j}) = \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{j}{k} \times \frac{1}{k^2}$$
$$= \frac{k \times k(k+1)}{2k^3} = \frac{k+1}{2k}.$$

Therefore, we find that

$$\lim_{k\to\infty} R(f, \mathbf{P}_k, A_k) = 0, \qquad \lim_{k\to\infty} R(f, \mathbf{P}_k, B_k) = \frac{1}{2}.$$

Since the two limits are not the same, we conclude that $f : \mathbf{I} \to \mathbb{R}$ is not Riemann integrable.

Now we return to the proof of Theorem 6.10. To prove this theorem, it is easier to show that (a) is equivalent to (c), and (b) is equivalent to (c). We will prove the equivalence of (a) and (c). The proof of the equivalence of (b) and (c) is left to the exercises. It is a consequence of the inequality

$$L(f, \mathbf{P}) \leq R(f, \mathbf{P}, A) \leq U(f, \mathbf{P}),$$

which holds for any partition $\mathbf{P}$ of the rectangle $\mathbf{I}$, and any choice of intermediate points A for the partition $\mathbf{P}$.

By Theorem 6.7, (a) is equivalent to

(a′) For every $\varepsilon > 0$, there is a partition $\mathbf{P}$ of $\mathbf{I}$ such that

$$U(f, \mathbf{P}) - L(f, \mathbf{P}) < \varepsilon.$$

Thus, to prove the equivalence of (a) and (c), it is sufficient to show the equivalence of (a′) and (c). But then (c) implies (a′) is obvious. Hence, we are left with the most technical part, which is the proof of (a′) implies (c).

We formulate this as a standalone theorem.

Theorem 6.13

Let $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$, and let $\mathbf{P}_0$ be a fixed a partition of $\mathbf{I}$. Given that $f : \mathbf{I} \to \mathbb{R}$ is a bounded function defined on $\mathbf{I}$, for any $\varepsilon > 0$, there is a $\delta > 0$ such that for all partitions $\mathbf{P}$ of $\mathbf{I}$, if $|\mathbf{P}| < \delta$, then

$$U(f, \mathbf{P}) - L(f, \mathbf{P}) < U(f, \mathbf{P}_0) - L(f, \mathbf{P}_0) + \varepsilon. \tag{6.1}$$

If Theorem 6.13 is proved, we can show that (a′) implies (c) in Theorem 6.10 as follows. Given $\varepsilon > 0$, (a′) implies that we can choose a $\mathbf{P}_0$ such that

$$U(f, \mathbf{P}_0) - L(f, \mathbf{P}_0) < \frac{\varepsilon}{2}.$$

By Theorem 6.13, there is a $\delta > 0$ such that for all partitions $\mathbf{P}$ of $\mathbf{I}$, if $|\mathbf{P}| < \delta$, then

$$U(f, \mathbf{P}) - L(f, \mathbf{P}) < U(f, \mathbf{P}_0) - L(f, \mathbf{P}_0) + \frac{\varepsilon}{2} < \varepsilon.$$

This proves that (a′) implies (c).

Hence, it remains for us to prove theorem 6.13. Let us introduce some additional notations. Given the rectangle $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$, for $1 \leq i \leq n$, let

$$\begin{aligned} \mathcal{S}_i &= \frac{\text{vol}\,(\mathbf{I})}{b_i - a_i} \\ &= (b_1 - a_1) \times \cdots \times (b_{i-1} - a_{i-1})(b_{i+1} - a_{i+1}) \times \cdots \times (b_n - a_n). \end{aligned} \tag{6.2}$$

This is the area of the bounday of $\mathbf{I}$ that is contained in the hyperplane $x_i = a_i$ or $x_i = b_i$. For example, when $n = 2$, $\mathbf{I} = [a_1, b_1] \times [a_2, b_2]$, $\mathcal{S}_1 = b_2 - a_2$ is the length of the vertical side, while $\mathcal{S}_2 = b_1 - a_1$ is the length of the horizontal side of the rectangle $\mathbf{I}$.

Proof of Theorem 6.13

Since $f : \mathbf{I} \to \mathbb{R}$ is bounded, there is a positive number M such that

$$|f(\mathbf{x})| \leq M \qquad \text{for all } \mathbf{x} \in \mathbf{I}.$$

Assume that $\mathbf{P}_0 = (\widetilde{P}_1, \ldots, \widetilde{P}_n)$. For $1 \leq i \leq n$, let k_i be the number of intervals in the partition $\widetilde{P}_i$. Let

$$K = \max\{k_1, \ldots, k_n\},$$

and

$$\mathcal{S} = \mathcal{S}_1 + \cdots + \mathcal{S}_n,$$

where $\mathcal{S}_i$, $1 \leq i \leq n$ are defined by (6.2). Given $\varepsilon > 0$, let

$$\delta = \frac{\varepsilon}{4MK\mathcal{S}}.$$

Then $\delta > 0$. If $\mathbf{P} = (P_1, \ldots, P_n)$ is a partition of $\mathbf{I}$ with $|\mathbf{P}| < \delta$, we want to show that (6.1) holds. Let $\mathbf{P}^* = (P_1^*, \ldots, P_n^*)$ be the common refinement of $\mathbf{P}_0$ and $\mathbf{P}$ such that P_i^* is the partition of $[a_i, b_i]$ that contains all the partition points of $\widetilde{P}_i$ and P_i. For $1 \leq i \leq n$, let U_i be the collection of intervals in P_i which contain partition points of $\widetilde{P}_i$, and let V_i be the collection of the intervals of P_i that is not in U_i. Each interval in V_i must be in the interior of one of the intervals in $\widetilde{P}_i$. Thus, each interval in V_i is an interval in the partition P_i^*. Since each partition point of $\widetilde{P}_i$ can be contained in at most two intervals of P_i, but the first and last partition points of P_i and $\widetilde{P}_i$ are the same, we find that $|U_i| \leq 2k_i$.

Since $|P_i| \leq |\mathbf{P}| < \delta$, each interval in P_i has length less than δ. Therefore, the sum of the lengths of the intervals in U_i is less than $2k_i\delta$. Let

$$\mathcal{Q}_i = \left\{ \mathbf{J} \in \mathcal{J}_{\mathbf{P}} \,|\, \text{the } i^{\text{th}}\text{-edge of } \mathbf{J} \text{ is from } U_i \right\}.$$

Then

$$\sum_{\mathbf{J} \in \mathcal{Q}_i} \text{vol}\,(\mathbf{J}) < 2k_i\delta\mathcal{S}_i \leq 2K\delta\mathcal{S}_i.$$

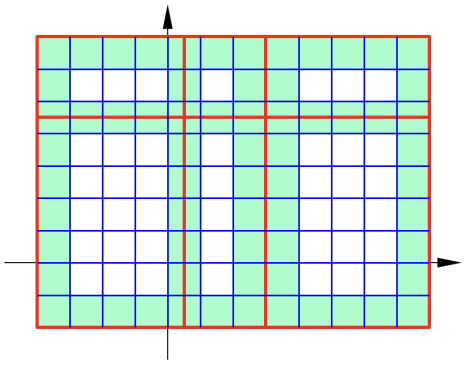


Figure 6.7: The partitions $\mathbf{P}_0$ and $\mathbf{P}$ in the proof of Theorem 6.13, $\mathbf{P}_0$ is the partition with red grids, while $\mathbf{P}$ is the partition with blue grids. Those shaded rectangles are rectangles in $\mathbf{P}$ that contain partition points of $\mathbf{P}_0$.

Now let

$$\mathcal{Q} = \bigcup_{i=1}^{n} \mathcal{Q}_i.$$

Then

$$\sum_{\mathbf{J}\in\mathcal{Q}} \text{vol}\,(\mathbf{J}) < 2K\delta \sum_{i=1}^{n} \mathcal{S}_i = 2K\delta\mathcal{S}.$$

For each of the rectangles $\mathbf{J}$ that is in $\mathcal{Q}$, we do a simple estimate

$$M_{\mathbf{J}} - m_{\mathbf{J}} \leq 2M.$$

Therefore,

$$\sum_{\mathbf{J}\in\mathcal{Q}} (M_{\mathbf{J}} - m_{\mathbf{J}})\,\text{vol}\,(\mathbf{J}) < 4MK\delta\mathcal{S} \leq \varepsilon.$$

For the rectangles $\mathbf{J}$ that are in $\mathcal{J}_{\mathbf{P}} \setminus \mathcal{Q}$, each of them is a rectangle in the partition $\mathbf{P}^*$. Therefore,

$$\sum_{\mathbf{J}\in\mathcal{J}_{\mathbf{P}}\setminus\mathcal{Q}} (M_{\mathbf{J}} - m_{\mathbf{J}})\,\text{vol}\,(\mathbf{J}) \leq U(f,\mathbf{P}^*) - L(f,\mathbf{P}^*) \leq U(f,\mathbf{P}_0) - L(f,\mathbf{P}_0).$$

Hence,

$$\begin{aligned} U(f,\mathbf{P}) - L(f,\mathbf{P}) &= \sum_{\mathbf{J}\in\mathcal{J}_\mathbf{P}} (M_\mathbf{J} - m_\mathbf{J})\,\mathrm{vol}\,(\mathbf{J}) \\ &= \sum_{\mathbf{J}\in\mathcal{J}_\mathbf{P}\setminus\mathcal{Q}} (M_\mathbf{J} - m_\mathbf{J})\,\mathrm{vol}\,(\mathbf{J}) + \sum_{\mathbf{J}\in\mathcal{Q}} (M_\mathbf{J} - m_\mathbf{J})\,\mathrm{vol}\,(\mathbf{J}) \\ &< U(f,\mathbf{P}_0) - L(f,\mathbf{P}_0) + \varepsilon. \end{aligned}$$

This completes the proof.

Finally we extend Riemann integrals to functions $f : \mathfrak{D} \to \mathbb{R}$ that are defined on bounded subsets $\mathfrak{D}$ of $\mathbb{R}^n$. If $\mathfrak{D}$ is bounded, there is a positive number L such that

$$\|\mathbf{x}\| \le L \qquad \text{for all } \mathbf{x} \in \mathfrak{D}.$$

This implies that $\mathfrak{D}$ is contained in the closed rectangle $\mathbf{I}_L = \prod_{i=1}^{n} [-L, L]$. To define the Riemann integral of $f : \mathfrak{D} \to \mathbb{R}$, we need to extend the domain of f from $\mathfrak{D}$ to $\mathbf{I}_L$. To avoid affecting the integral, we should extend by zero.

Definition 6.11 Zero Extension

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$, and let $f : \mathfrak{D} \to \mathbb{R}$ be a function defined on $\mathfrak{D}$. The zero extension of $f : \mathfrak{D} \to \mathbb{R}$ is the function $\check{f} : \mathbb{R}^n \to \mathbb{R}$ which is defined as

$$\check{f}(\mathbf{x}) = \begin{cases} f(\mathbf{x}), & \text{if } \mathbf{x} \in \mathfrak{D}, \\ 0, & \text{if } \mathbf{x} \notin \mathfrak{D}. \end{cases}$$

If $\mathcal{U}$ is any subset of $\mathbb{R}^n$ that contains $\mathfrak{D}$, then the zero extension of f to $\mathcal{U}$ is the function $\check{f} : \mathcal{U} \to \mathbb{R}$.

Obviously, if $f : \mathfrak{D} \to \mathbb{R}$ is a bounded function, its zero extension $\check{f} : \mathbb{R}^n \to \mathbb{R}$ is also bounded. Since we have defined Riemann integrability for a bounded function $g : \mathbf{I} \to \mathbb{R}$ that is defined on a closed rectangle $\mathbf{I}$, it is natural to say that a function $f : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable if its zero extension $\check{f} : \mathbf{I} \to \mathbb{R}$ to a closed rectangle $\mathbf{I}$ is Riemann integrable, and define

$$\int_{\mathfrak{D}} f = \int_{\mathbf{I}} \check{f}.$$

For this to be unambiguous, we have to check that if $\mathbf{I}_1$ and $\mathbf{I}_2$ are closed rectangles that contain the bounded set $\mathfrak{D}$, the zero extension $\check{f} : \mathbf{I}_1 \to \mathbb{R}$ is Riemann integrable if and only if the zero extension $\check{f} : \mathbf{I}_2 \to \mathbb{R}$ is Riemann integrable. Moreover,

$$\int_{\mathbf{I}_1} \check{f} = \int_{\mathbf{I}_2} \check{f}.$$

This small technicality would be proved in Section 6.2. Assuming this, we can give the following formal definition for Riemann integrality of a bounded function defined on a bounded domain.

Definition 6.12 Riemann Integrals of General Functions

Let $\mathfrak{D}$ be a bounded subset of $\mathbb{R}^n$, and let $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$ be a closed rectangle in $\mathbb{R}^n$ that contains $\mathfrak{D}$. Given that $f : \mathfrak{D} \to \mathbb{R}$ is a bounded function defined on $\mathfrak{D}$, we say that $f : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable if its zero extension $\check{f} : \mathbf{I} \to \mathbb{R}$ is Riemann integrable. If this is the case, we define the integral of f over $\mathfrak{D}$ as

$$\int_{\mathfrak{D}} f = \int_{\mathbf{I}} \check{f}.$$

Example 6.17

Let $\mathbf{I} = [0, 1] \times [0, 1]$, and let $f : \mathbf{I} \to \mathbb{R}$ be the function defined as

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

which is considered in Example 6.14. Let

$$\mathfrak{D} = \{(x, y) \in \mathbf{I} \,|\, y \leq x\},$$

and let $g : \mathfrak{D} \to \mathbb{R}$ be the constant function $g(\mathbf{x}) = 1$. Then $f : \mathbf{I} \to \mathbb{R}$ is the zero extension of g to the square $\mathbf{I}$ that contains $\mathfrak{D}$.

In Example 6.14, we have shown that $f : \mathbf{I} \to \mathbb{R}$ is Riemann integrable and

$$\int_{\mathbf{I}} f(\mathbf{x}) d\mathbf{x} = \frac{1}{2}.$$

Therefore, $g : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable and

$$\int_{\mathfrak{D}} g(\mathbf{x}) d\mathbf{x} = \frac{1}{2}.$$

Remark 6.1

Here we make two remarks about the Riemann integrals.

1. When $f : \mathfrak{D} \to \mathbb{R}$ is the constant function, we should expect that it is Riemann integrable if and only if $\mathfrak{D}$ has a volume, which should be defined as
$$\text{vol}\,(\mathfrak{D}) = \int_{\mathfrak{D}} d\mathbf{x}.$$
2. If $f : \mathfrak{D} \to \mathbb{R}$ is a nonnegative continuous function defined on the bounded set $\mathfrak{D}$ that has a volume, we would expect that $f : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable, and the integral $\int_{\mathfrak{D}} f(\mathbf{x}) d\mathbf{x}$ gives the volume of the solid bounded between $\mathfrak{D}$ and the graph of f.

In Section 6.3, we will give a characterization of sets $\mathfrak{D}$ that have volumes. We will also prove that if $f : \mathfrak{D} \to \mathbb{R}$ is a continuous function defined on a set $\mathfrak{D}$ that has volume, then $f : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable.

Exercises 6.1

Question 1

Let $\mathbf{I} = [-5, 8] \times [2, 5]$, and let $\mathbf{P} = (P_1, P_2)$ be the partition of $\mathbf{I}$ with $P_1 = \{-5, -1, 2, 7, 8\}$ and $P_2 = \{2, 4, 5\}$. Find gap of the partition $\mathbf{P}$.

Question 2

Let $\mathbf{I} = [-5, 8] \times [2, 5]$, and let $f : \mathbf{I} \to \mathbb{R}$ be the function defined as $f(x, y) = x^2 + 2y$. Consider the partition $\mathbf{P} = (P_1, P_2)$ of $\mathbf{I}$ with $P_1 = \{-5, -1, 2, 7, 8\}$ and $P_2 = \{2, 4, 5\}$. Find the Darboux lower sum $L(f, \mathbf{P})$ and the Darboux upper sum $U(f, \mathbf{P})$.

Question 3

Let $\mathbf{I} = [-5, 8] \times [2, 5]$, and let $f : \mathbf{I} \to \mathbb{R}$ be the function defined as $f(x, y) = x^2 + 2y$. Consider the partition $\mathbf{P} = (P_1, P_2)$ of $\mathbf{I}$ with $P_1 = \{-5, -1, 2, 7, 8\}$ and $P_2 = \{2, 4, 5\}$. For each rectangle $\mathbf{J} = [a, b] \times [c, d]$ in the partition $\mathbf{P}$, let $\boldsymbol{\alpha}_\mathbf{J} = (a, c)$ and $\boldsymbol{\beta}_\mathbf{J} = (b, d)$. Find the Riemann sums $R(f, \mathbf{P}, A)$ and $R(f, \mathbf{P}, B)$, where $A = \{\boldsymbol{\alpha}_\mathbf{J}\}$ and $B = \{\boldsymbol{\beta}_\mathbf{J}\}$.

Question 4

Let $\mathbf{I} = [-1, 1] \times [2, 5]$, and let $f : \mathbf{I} \to \mathbb{R}$ be the function defined as

$$f(x, y) = \begin{cases} 1, & \text{if } x \text{ and } y \text{ are rational,} \\ 0, & \text{otherwise.} \end{cases}$$

(a) Given that $\mathbf{P}$ is a partition of $\mathbf{I}$, find the Darboux lower sum $L(f, \mathbf{P})$ and the Darboux upper sum $U(f, \mathbf{P})$.

(b) Find the lower integral $\underline{\int_\mathbf{I}} f$ and the upper integral $\overline{\int_\mathbf{I}} f$.

(c) Explain why $f : \mathbf{I} \to \mathbb{R}$ is not Riemann integrable.

Question 5

Let $\mathbf{I} = [0, 4] \times [0, 2]$. Consider the function $f : \mathbf{I} \to \mathbb{R}$ defined as

$$f(x, y) = 2x + 3y + 1.$$

For $k \in \mathbb{Z}^+$, let $\mathbf{P}_k$ be the uniformly regular partition of $\mathbf{I} = [0, 4] \times [0, 2]$ into k^2 rectangles.

(a) For each $k \in \mathbb{Z}^+$, compute the Darboux lower sum $L(f, \mathbf{P}_k)$ and the Darboux upper sum $U(f, \mathbf{P}_k)$.

(b) Show that $f : \mathbf{I} \to \mathbb{R}$ is Riemann integrable and find the integral $\displaystyle\int_{\mathbf{I}} f$.

Question 6

Let $\mathbf{I} = \displaystyle\prod_{i=1}^{n} [a_i, b_i]$, and let $f : \mathbf{I} \to \mathbb{R}$ be a function defined on $\mathbf{I}$. Show that the following are equivalent.

(a) There exists a number I that satisfies the following. For any $\varepsilon > 0$, there exists a $\delta > 0$ such that if $\mathbf{P}$ is a partition of the rectangle $\mathbf{I}$ with $|\mathbf{P}| < \delta$, then

$$|R(f, \mathbf{P}, A) - I| < \varepsilon$$

for any choice of intermediate points $A = \{\boldsymbol{\xi}_\mathbf{J}\}$ for the partition $\mathbf{P}$.

(b) For any $\varepsilon > 0$, there exists a $\delta > 0$ such that if $\mathbf{P}$ is a partition of the rectangle $\mathbf{I}$ with $|\mathbf{P}| < \delta$, then

$$U(f, \mathbf{P}) - L(f, \mathbf{P}) < \varepsilon.$$

6.2 Properties of Riemann Integrals

In this section, we discuss properties of Riemann integrals. Let us first consider Riemann integrals of functions $f : \mathbf{I} \to \mathbb{R}$ defined on closed rectangles of the form $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$. Using some of these properties, we prove that the definition of Riemann integrabililty for functions $f : \mathfrak{D} \to \mathbb{R}$ defined on general bounded sets, as given in Section 6.1, is unambiguous. Finally, we will extend the properties of Riemann integrals to functions $f : \mathfrak{D} \to \mathbb{R}$ defined on bounded sets.

Linearity is one of the most important properties. For functions defined on closed rectangles of the form $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$, the proof is straightforward using the Riemann sum definition of Riemann integrability, as in the single variable case.

Theorem 6.14 Linearity

Let $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$, and let $f : \mathbf{I} \to \mathbb{R}$ and $g : \mathbf{I} \to \mathbb{R}$ be Riemann integrable functions. For any real numbers α and β, $(\alpha f + \beta g) : \mathbf{I} \to \mathbb{R}$ is also Riemann integrable, and

$$\int_{\mathbf{I}} (\alpha f + \beta g) = \alpha \int_{\mathbf{I}} f + \beta \int_{\mathbf{I}} g.$$

Sketch of Proof

If $\mathbf{P}$ is a partition of $\mathbf{I}$ and A is a set of intermediate points for $\mathbf{P}$, then

$$R(\alpha f + \beta g, \mathbf{P}, A) = \alpha R(f, \mathbf{P}, A) + \beta R(g, \mathbf{P}, A).$$

The results follows by taking the $|\mathbf{P}| \to 0$ limit.

Example 6.18

Let $\mathbf{I} = [0, 2] \times [0, 2]$, and let $f : \mathbf{I} \to \mathbb{R}$ and $g : \mathbf{I} \to \mathbb{R}$ be Riemann integrable functions. Find the integrals $\int_{\mathbf{I}} f$ and $\int_{\mathbf{I}} g$ if

$$f(x, y) = g(y, x) \quad \text{and} \quad (f + g)(x, y) = 6 \qquad \text{for all } (x, y) \in \mathbf{I}.$$

Solution

Since $\mathbf{I}$ is symmetric with respect to the line $y = x$ and $f(x, y) = g(y, x)$ for all $(x, y) \in \mathbf{I}$, we have $\int_{\mathbf{I}} f = \int_{\mathbf{I}} g$. By linearity,

$$\int_{\mathbf{I}} f + \int_{\mathbf{I}} g = \int_{\mathbf{I}} (f + g) = 6 \times \text{vol}\,(\mathbf{I}) = 24.$$

Hence,

$$\int_{\mathbf{I}} f = \int_{\mathbf{I}} g = 12.$$

The following theorem is about the integral of a nonnegative function.

Theorem 6.15

Let $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$, and let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function defined on $\mathbf{I}$. Assume that $f(\mathbf{x}) \geq 0$ for all $\mathbf{x}$ in $\mathbf{I}$. If $f : \mathbf{I} \to \mathbb{R}$ is Riemann integrable, then

$$\int_{\mathbf{I}} f \;\geq\; 0.$$

Proof

For any partition $\mathbf{P}$ of $\mathbf{I}$, $L(f, \mathbf{P}) \geq 0$. Therefore,

$$\int_{\mathbf{I}} f \;=\; \underline{\int_{\mathbf{I}}} f \;\geq\; L(f, \mathbf{P}) \;\geq\; 0.$$

The monotonicity theorem then follows from linearity and Theorem 6.15.

Theorem 6.16 Monotonicity

Let $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$, and let $f : \mathbf{I} \to \mathbb{R}$ and $g : \mathbf{I} \to \mathbb{R}$ be Riemann integrable functions. If $f(\mathbf{x}) \geq g(\mathbf{x})$ for all $\mathbf{x}$ in $\mathbf{I}$, then

$$\int_{\mathbf{I}} f \;\geq\; \int_{\mathbf{I}} g.$$

Proof

By linearity, the function $(f-g):\mathbf{I}\to\mathbb{R}$ is integrable, and

$$\int_{\mathbf{I}}(f-g)=\int_{\mathbf{I}}f-\int_{\mathbf{I}}g.$$

By Theorem 6.15, $\displaystyle\int_{\mathbf{I}}(f-g)\geq 0$, and the assertion follows.

The next important property is the additivity of the Riemann integrals.

Theorem 6.17 Additivity

Let $\mathbf{I}=\prod_{i=1}^{n}[a_i,b_i]$, and let $\mathbf{P}_0$ be a partition of $\mathbf{I}$. If $f:\mathbf{I}\to\mathbb{R}$ is a bounded function defined on $\mathbf{I}$, then $f:\mathbf{I}\to\mathbb{R}$ is Riemann integrable if and only if for each $\mathbf{J}\in\mathcal{J}_{\mathbf{P}_0}$, $f:\mathbf{J}\to\mathbb{R}$ is Riemann integrable. In such case, we also have

$$\int_{\mathbf{I}}f=\sum_{\mathbf{J}\in\mathcal{J}_{\mathbf{P}_0}}\int_{\mathbf{J}}f.$$

Proof

It is sufficient to consider the case that $\mathbf{P}_0=(P_1,\ldots,P_n)$ divides $\mathbf{I}$ into two rectangles $\mathbf{I}_1$ and $\mathbf{I}_2$ by having a partition point c inside the j^{th}-edge $[a_j,b_j]$ for some $1\leq j\leq n$. Namely, $P_j=\{a_j,c,b_j\}$, and for $i\neq j$, $P_i=\{a_i,b_i\}$. The general case can be proved by induction, adding one partition point at a time.

Assume that $f:\mathbf{I}\to\mathbb{R}$ is Riemann integrable. Given $\varepsilon>0$, there is a partition $\mathbf{P}$ of $\mathbf{I}$ such that

$$U(f,\mathbf{P})-L(f,\mathbf{P})<\varepsilon.$$

Let $\mathbf{P}^*$ be a common refinement of $\mathbf{P}$ and $\mathbf{P}_0$. Then

$$U(f,\mathbf{P}^*)-L(f,\mathbf{P}^*)\leq U(f,\mathbf{P})-L(f,\mathbf{P})<\varepsilon.$$

But $\mathbf{P}^*$ induces a partition $\mathbf{P}^*(\mathbf{I}_1)$ and $\mathbf{P}^*(\mathbf{I}_2)$ of $\mathbf{I}_1$ and $\mathbf{I}_2$, and we have

$$U(f,\mathbf{P}^*) = U(f,\mathbf{P}^*(\mathbf{I}_1)) + U(f,\mathbf{P}^*(\mathbf{I}_2)),$$

$$L(f,\mathbf{P}^*) = L(f,\mathbf{P}^*(\mathbf{I}_1)) + L(f,\mathbf{P}^*(\mathbf{I}_2)).$$

Therefore,

$$U(f,\mathbf{P}^*(\mathbf{I}_1)) - L(f,\mathbf{P}^*(\mathbf{I}_1)) + U(f,\mathbf{P}^*(\mathbf{I}_2)) - L(f,\mathbf{P}^*(\mathbf{I}_2)) < \varepsilon.$$

This implies that

$$U(f,\mathbf{P}^*(\mathbf{I}_j)) - L(f,\mathbf{P}^*(\mathbf{I}_j)) < \varepsilon \qquad \text{for } j = 1, 2.$$

Hence, $f : \mathbf{I}_1 \to \mathbb{R}$ and $f : \mathbf{I}_2 \to \mathbb{R}$ are Riemann integrable.
Conversely, assume that $f : \mathbf{I}_1 \to \mathbb{R}$ and $f : \mathbf{I}_2 \to \mathbb{R}$ are Riemann integrable. Let $\{\mathbf{P}_{1,k}\}$ and $\{\mathbf{P}_{2,k}\}$ be Archimedes sequences of partitions for $f : \mathbf{I}_1 \to \mathbb{R}$ and $f : \mathbf{I}_2 \to \mathbb{R}$ respectively. Then

$$\int_{\mathbf{I}_j} f = \lim_{k\to\infty} U(f,\mathbf{P}_{j,k}) = \lim_{k\to\infty} L(f,\mathbf{P}_{j,k}) \qquad \text{for } j = 1, 2.$$

For $k \in \mathbb{Z}^+$, let $\mathbf{P}_k^*$ be the partition of $\mathbf{I}$ obtained by taking unions of partition points in $\mathbf{P}_{1,k}$ and $\mathbf{P}_{2,k}$. Then $\mathbf{P}_{1,k} = \mathbf{P}_k^*(\mathbf{I}_1)$ and $\mathbf{P}_{2,k} = \mathbf{P}_k^*(\mathbf{I}_2)$. It follows that

$$U(f,\mathbf{P}_k^*) = U(f,\mathbf{P}_{1,k}) + U(f,\mathbf{P}_{2,k}), \quad L(f,\mathbf{P}_k^*) = L(f,\mathbf{P}_{1,k}) + L(f,\mathbf{P}_{2,k}).$$

Therefore,

$$\lim_{k\to\infty} (U(f,\mathbf{P}_k^*) - L(f,\mathbf{P}_k^*)) = 0.$$

Hence, $\{\mathbf{P}_k^*\}$ is an Archimedes sequence of partitions for $f : \mathbf{I} \to \mathbb{R}$. This shows that $f : \mathbf{I} \to \mathbb{R}$ is Riemann integrable, and

$$\int_{\mathbf{I}} f = \lim_{k\to\infty} U(f,\mathbf{P}_k^*) = \lim_{k\to\infty} (U(f,\mathbf{P}_{1,k}) + U(f,\mathbf{P}_{2,k})) = \int_{\mathbf{I}_1} f + \int_{\mathbf{I}_2} f.$$

Next we state a lemma which is useful.

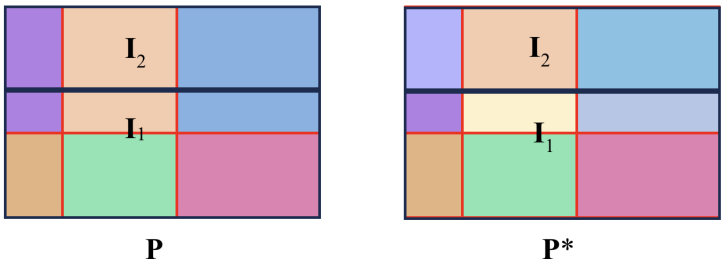


Figure 6.8: A partition $\mathbf{P}$ of $\mathbf{I}$ and the refined partition $\mathbf{P}^*$ that induces partitions on $\mathbf{I}_1$ and $\mathbf{I}_2$.

Lemma 6.18

Given that $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$ is a closed rectangle in $\mathbb{R}^n$, let

$$\omega = \frac{1}{2} \min\{b_i - a_i \,|\, 1 \leq i \leq n\}.$$

(i) Given $\eta > 0$, let $\mathbf{I}_\eta$ be the closed rectangle $\mathbf{I}_\eta = \prod_{i=1}^{n} [a_i - \eta, b_i + \eta]$. For any $\varepsilon > 0$, there exists $\delta > 0$ such that for any $0 < \eta < \delta$, $0 < \text{vol}\,(\mathbf{I}_\eta) - \text{vol}\,(\mathbf{I}) < \varepsilon$.

(ii) Given $0 < \kappa < \omega$, let $\check{\mathbf{I}}_\kappa$ be the closed rectangle $\check{\mathbf{I}}_\kappa = \prod_{i=1}^{n} [a_i + \kappa, b_i - \kappa]$. For any $\varepsilon > 0$, there exists $0 < \delta \leq \omega$ such that for any $0 < \kappa < \delta$, $0 < \text{vol}\,(\mathbf{I}) - \text{vol}\,(\check{\mathbf{I}}_\kappa) < \varepsilon$.

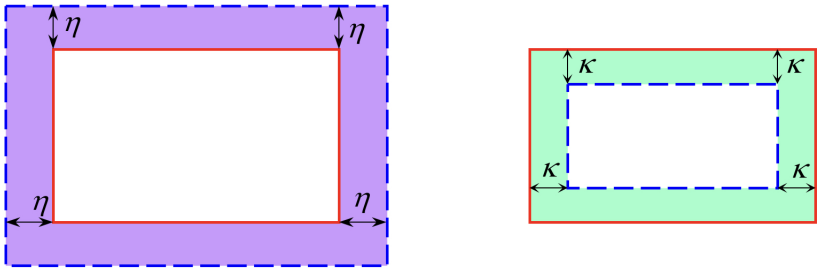


Figure 6.9: Enlarging or shrinking a rectangle by an arbitrary amount.

This lemma says that one can enlarge or shrink a rectangle by an arbitrarily small amount. It can be proved by elementary means. But here we use some analysis technique to prove it.

Proof

We prove part (i). The argument for part (ii) is the same. Consider the function $h : [0, \infty) \to \mathbb{R}$ defined by

$$h(\eta) = \text{vol}\,(\mathbf{I}_\eta) = \prod_{i=1}^{n}(b_i - a_i + 2\eta).$$

As a function of η, $h(\eta)$ is a polynomial, and it is a strictly increasing continuous function. The assertion is basically the definition of the limit $\lim_{\eta \to 0^+} h(\eta) = h(0)$.

The following theorem says that a bounded function $f : \mathbf{I} \to \mathbb{R}$ which is identically zero on the interior of $\mathbf{I}$ is Riemann integrable with integral 0. This is something we would have expected.

Theorem 6.19

Let $\mathbf{I} = \prod_{i=1}^{n}[a_i, b_i]$, and let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function such that

$$f(\mathbf{x}) = 0 \qquad \text{for all } \mathbf{x} \in \text{int}\,(\mathbf{I}).$$

Then $f : \mathbf{I} \to \mathbb{R}$ is Riemann integrable and

$$\int_{\mathbf{I}} f = 0.$$

Proof

Let $\omega = \frac{1}{2}\min\{b_i - a_i \,|\, 1 \le i \le n\}$. Since $f : \mathbf{I} \to \mathbb{R}$ is bounded, there is a positive number M such that

$$|f(\mathbf{x})| \le M \qquad \text{for all } \mathbf{x} \in \mathbf{I}.$$

By Lemma 6.18, there is a $\kappa \in (0, \omega)$ such that

$$\text{vol}\,(\mathbf{I}) - \text{vol}\,(\mathbf{I}_\kappa) < \frac{\varepsilon}{M},$$

where $\mathbf{I}_\kappa = \prod_{i=1}^{n} [a_i + \kappa, b_i - \kappa]$. It is a rectangle that is contained in int $(\mathbf{I})$. Let $\mathbf{P} = (P_1, \ldots, P_n)$ be the partition of $\mathbf{I}$ with $P_i = \{a_i, a_i + \kappa, b_i - \kappa, b_i\}$. Then $\mathbf{I}_\kappa$ is one of the rectangles in the partition $\mathbf{P}$. On $\mathbf{J} = \mathbf{I}_\kappa$, $f(\mathbf{x}) = 0$, and so $M_\mathbf{J} = m_\mathbf{J} = 0$. For all other rectangles $\mathbf{J}$ in $\mathcal{J}_\mathbf{P}$, we use the crude estimate

$$-M \leq m_\mathbf{J} \leq M_\mathbf{J} \leq M.$$

Then

$$\begin{aligned} L(f, \mathbf{P}) &= \sum_{\mathbf{J} \in \mathcal{J}_\mathbf{P}} m_\mathbf{J}\, \text{vol}\,(\mathbf{J}) = \sum_{\mathbf{J} \in \mathcal{J}_\mathbf{P} \setminus \{\mathbf{I}_\kappa\}} m_\mathbf{J}\, \text{vol}\,(\mathbf{J}) \\ &\geq -M\,(\text{vol}\,(\mathbf{I}) - \text{vol}\,(\mathbf{I}_\kappa)) > -\varepsilon. \end{aligned}$$

In the same way, we find that $U(f, \mathbf{P}) < \varepsilon$. Since $\varepsilon > 0$ is arbitrary, we find that

$$\underline{\int_\mathbf{I}} f \geq 0 \quad \text{and} \quad \overline{\int_\mathbf{I}} f \leq 0.$$

Since $\underline{\int_\mathbf{I}} f \leq \overline{\int_\mathbf{I}} f$, we conclude that

$$\underline{\int_\mathbf{I}} f = \overline{\int_\mathbf{I}} f = 0.$$

This proves that $f : \mathbf{I} \to \mathbb{R}$ is Riemann integrable and

$$\int_\mathbf{I} f = 0.$$

Now let us give a proof that the definition given in Section 6.1 for a bounded function $f : \mathfrak{D} \to \mathbb{R}$ defined on a bounded subset of $\mathbb{R}^n$ to be Riemann integrable is unambiguous. The crucial point is the following.

Lemma 6.20

Let $\mathbf{I} = \prod_{i=1}^{n}[a_i, b_i]$ and $\check{\mathbf{I}} = \prod_{i=1}^{n}[\check{a}_i, \check{b}_i]$ be closed rectangles in $\mathbb{R}^n$ such that $\mathbf{I} \subset \check{\mathbf{I}}$, and let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function defined on $\mathbf{I}$. Then $f : \mathbf{I} \to \mathbb{R}$ is Riemann integrable if and only if its zero extension $\check{f} : \check{\mathbf{I}} \to \mathbb{R}$ is Riemann integrable. In such case, we also have

$$\int_{\mathbf{I}} f = \int_{\check{\mathbf{I}}} \check{f}.$$

Proof

Let $\check{\mathbf{P}} = \{\check{P}_1, \ldots, \check{P}_n\}$ be the partition of $\check{\mathbf{I}}$ such that the set $\check{P}_i$ is the set that contains $\check{a}_i, a_i, b_i, \check{b}_i$.

For each rectangle $\mathbf{J}$ in $\mathcal{J}_{\check{\mathbf{P}}} \setminus \{\mathbf{I}\}$, it is disjoint from the interior of $\mathbf{I}$. Hence, $\check{f}$ vanishes in the interior of $\mathbf{J}$. By Theorem 6.19, $\check{f} : \mathbf{J} \to \mathbb{R}$ is Riemann integrable and $\int_{\mathbf{J}} \check{f} = 0$. It follows from the additivity theorem that $\check{f} : \check{\mathbf{I}} \to \mathbb{R}$ is Riemann integrable if and only if $\check{f} : \mathbf{I} \to \mathbb{R}$ is Riemann integrable, and

$$\int_{\check{\mathbf{I}}} \check{f} = \int_{\mathbf{I}} \check{f}.$$

However, restricted to $\mathbf{I}$, $\check{f}(\mathbf{x}) = f(\mathbf{x})$. Hence, $\check{f} : \check{\mathbf{I}} \to \mathbb{R}$ is Riemann integrable if and only if $f : \mathbf{I} \to \mathbb{R}$ is Riemann integrable. In such case, we have

$$\int_{\check{\mathbf{I}}} \check{f} = \int_{\mathbf{I}} f.$$

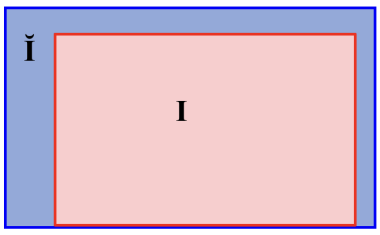


Figure 6.10: The rectangle $\mathbf{I}$ is contained in the rectangle $\check{\mathbf{I}}$.

Finally we can prove the main result.

Theorem 6.21

Let $\mathfrak{D}$ be a bounded set in $\mathbb{R}^n$, and let $f : \mathfrak{D} \to \mathbb{R}$ be a bounded function defined on $\mathfrak{D}$. The definition for Riemann integrability of $f : \mathfrak{D} \to \mathbb{R}$ is unambiguous. Namely, if $\mathbf{I}_1 = \prod_{i=1}^{n} [a_i', b_i']$ and $\mathbf{I}_2 = \prod_{i=1}^{n} [a_i'', b_i'']$ contain $\mathfrak{D}$, the zero extension $\check{f} : \mathbf{I}_1 \to \mathbb{R}$ is Riemann integrable if and only if the zero extension $\check{f} : \mathbf{I}_2 \to \mathbb{R}$ is Riemann integrable. In the latter case,

$$\int_{\mathbf{I}_1} \check{f} = \int_{\mathbf{I}_2} \check{f},$$

and so we can define unambiguously

$$\int_{\mathfrak{D}} f = \int_{\mathbf{I}} \check{f},$$

where $\mathbf{I}$ is any rectangle of the form $\prod_{i=1}^{n} [a_i, b_i]$ that contains $\mathfrak{D}$.

Proof

Let $\mathbf{I} = \mathbf{I}_1 \cap \mathbf{I}_2$. Then $\mathbf{I}$ is a rectangle that is contained in $\mathbf{I}_1$ and $\mathbf{I}_2$. Lemma 6.20 then says that $\check{f} : \mathbf{I}_1 \to \mathbb{R}$ is Riemann integrable if and only if $\check{f} : \mathbf{I} \to \mathbb{R}$ is Riemann integrable, if and only if $\check{f} : \mathbf{I}_2 \to \mathbb{R}$ is Riemann integrable. In latter case,

$$\int_{\mathbf{I}_1} \check{f} = \int_{\mathbf{I}} \check{f} = \int_{\mathbf{I}_2} \check{f}.$$

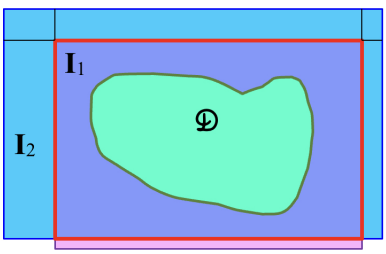


Figure 6.11: The set $\mathfrak{D}$ is contained in the rectangles $\mathbf{I}_1$ and $\mathbf{I}_2$.

Now we can extend the linearity and monotonicity to Riemann integrals over any bounded domains.

Theorem 6.22 Linearity

Let $\mathfrak{D}$ be bounded subset of $\mathbb{R}^n$, and let $f : \mathfrak{D} \to \mathbb{R}$ and $g : \mathfrak{D} \to \mathbb{R}$ be Riemann integrable functions. For any real numbers α and β, $(\alpha f + \beta g) : \mathfrak{D} \to \mathbb{R}$ is also Riemann integrable, and

$$\int_{\mathfrak{D}} (\alpha f + \beta g) = \alpha \int_{\mathfrak{D}} f + \beta \int_{\mathfrak{D}} g.$$

Proof

Let $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$ be a closed rectangle that contains $\mathfrak{D}$, and let $\check{f} : \mathbf{I} \to \mathbb{R}$ and $\check{g} : \mathbf{I} \to \mathbb{R}$ be the zero extensions of $f : \mathfrak{D} \to \mathbb{R}$ and $g : \mathfrak{D} \to \mathbb{R}$ to $\mathbf{I}$. It is easy to check that $(\alpha \check{f} + \beta \check{g}) : \mathbf{I} \to \mathbb{R}$ is the zero extension of $(\alpha f + \beta g) : \mathfrak{D} \to \mathbb{R}$ to $\mathbf{I}$. Since $f : \mathfrak{D} \to \mathbb{R}$ and $g : \mathfrak{D} \to \mathbb{R}$ are Riemann integrable, $\check{f} : \mathbf{I} \to \mathbb{R}$ and $\check{g} : \mathbf{I} \to \mathbb{R}$ are Riemann integrable and

$$\int_{\mathbf{I}} \check{f} = \int_{\mathfrak{D}} f, \qquad \int_{\mathbf{I}} \check{g} = \int_{\mathfrak{D}} g.$$

By Theorem 6.14, $(\alpha \check{f} + \beta \check{g}) : \mathbf{I} \to \mathbb{R}$ is Riemann integrable, and

$$\int_{\mathbf{I}} (\alpha \check{f} + \beta \check{g}) = \alpha \int_{\mathbf{I}} \check{f} + \beta \int_{\mathbf{I}} \check{g} = \alpha \int_{\mathfrak{D}} f + \beta \int_{\mathfrak{D}} g.$$

It follows that $(\alpha f + \beta g) : \mathfrak{D} \to \mathbb{R}$ is also Riemann integrable, and

$$\int_{\mathfrak{D}} (\alpha f + \beta g) = \int_{\mathbf{I}} (\alpha \check{f} + \beta \check{g}) = \alpha \int_{\mathfrak{D}} f + \beta \int_{\mathfrak{D}} g.$$

Theorem 6.23

Let $\mathfrak{D}$ be a bounded subset of $\mathbb{R}^n$, and let $f : \mathfrak{D} \to \mathbb{R}$ be a bounded function defined on $\mathfrak{D}$. Assume that $f(\mathbf{x}) \geq 0$ for all $\mathbf{x}$ in $\mathfrak{D}$. If $f : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable, then

$$\int_{\mathfrak{D}} f \geq 0.$$

Proof

Let $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$ be a closed rectangle that contains $\mathfrak{D}$, and let $\check{f} : \mathbf{I} \to \mathbb{R}$ be the zero extension of $f : \mathfrak{D} \to \mathbb{R}$ to $\mathbf{I}$. Since $f : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable, $\check{f} : \mathbf{I} \to \mathbb{R}$ is also Riemann integrable. It is easy to check that $\check{f}(\mathbf{x}) \geq 0$ for all $\mathbf{x}$ in $\mathbf{I}$. Therefore,

$$\int_{\mathfrak{D}} f = \int_{\mathbf{I}} \check{f} \geq 0.$$

As before, monotonicity is a consequence of linearity and Theorem 6.23.

Theorem 6.24 Monotonicity

Let $\mathfrak{D}$ be a bounded subset of $\mathbb{R}^n$, and let $f : \mathfrak{D} \to \mathbb{R}$ and $g : \mathfrak{D} \to \mathbb{R}$ be Riemann integrable functions. If $f(\mathbf{x}) \geq g(\mathbf{x})$ for all $\mathbf{x}$ in $\mathfrak{D}$, then

$$\int_{\mathfrak{D}} f \geq \int_{\mathfrak{D}} g.$$

At the end of this section, we want to present two theorems whose proofs are almost verbatim those for the $n = 1$ case. The first theorem says that if a function is Riemann integrable, so is its absolute value.

Theorem 6.25 Absolute Value of Riemann Integrable Functions

Let $\mathfrak{D}$ be a bounded subset of $\mathbb{R}^n$, and let $f : \mathfrak{D} \to \mathbb{R}$ be a bounded function defined on $\mathfrak{D}$. If the function $f : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable, then the function $|f| : \mathfrak{D} \to \mathbb{R}$ is also Riemann integrable.

Sketch of Proof

If $\check{f} : \mathbf{I} \to \mathbb{R}$ is the zero extension of $f : \mathfrak{D} \to \mathbb{R}$ to the closed rectangle $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$ that contains $\mathfrak{D}$, then $|\check{f}| : \mathbf{I} \to \mathbb{R}$ is the zero extension of $|f| : \mathfrak{D} \to \mathbb{R}$. Hence, it is sufficient to consider the case where $\mathfrak{D}$ is a closed rectangle of the form $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$. The proof is almost the same as the $n = 1$ case. The key of the proof is the fact that for any subset A of $\mathbf{I}$,

$$\sup_{\mathbf{x}\in A} |f(\mathbf{x})| - \inf_{\mathbf{x}\in A} |f(\mathbf{x})| \le \sup_{\mathbf{x}\in A} f(\mathbf{x}) - \inf_{\mathbf{x}\in A} f(\mathbf{x}).$$

The second theorem says that products of Riemann integrable functions are Riemann integrable.

Theorem 6.26 Products of Riemann Integrable Functions

Let $\mathfrak{D}$ be a bounded subset of $\mathbb{R}^n$, and let $f : \mathfrak{D} \to \mathbb{R}$ and $g : \mathfrak{D} \to \mathbb{R}$ be bounded functions defined on $\mathfrak{D}$. If the functions $f : \mathfrak{D} \to \mathbb{R}$ and $g : \mathfrak{D} \to \mathbb{R}$ are Riemann integrable, then the function $(fg) : \mathfrak{D} \to \mathbb{R}$ is also Riemann integrable.

Sketch of Proof

It is sufficient to consider the case where $\mathfrak{D}$ is a closed rectangle of the form $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$. The proof is almost the same as the $n = 1$ case. The key of the proof is the fact that if M is positive number such that

$$|f(\mathbf{x})| \le M \quad \text{and} \quad |g(\mathbf{x})| \le M \qquad \text{for all } \mathbf{x} \in \mathbf{I},$$

then for any subset A of $\mathbf{I}$,

$$\begin{aligned} &\sup_{\mathbf{x}\in A} (fg)(\mathbf{x}) - \inf_{\mathbf{x}\in A} (fg)(\mathbf{x}) \\ &\le M \left(\sup_{\mathbf{x}\in A} f(\mathbf{x}) - \inf_{\mathbf{x}\in A} f(\mathbf{x}) + \sup_{\mathbf{x}\in A} g(\mathbf{x}) - \inf_{\mathbf{x}\in A} g(\mathbf{x}) \right). \end{aligned}$$

Exercises 6.2

Question 1

Let $\mathbf{I} = [0,3] \times [0,3]$, and let $f : \mathbf{I} \to \mathbb{R}$ and $g : \mathbf{I} \to \mathbb{R}$ be Riemann integrable functions. Suppose that

$$f(x,y) = g(y,x) \quad \text{and} \quad (3f + 2g)(x,y) = 10 \qquad \text{for all } (x,y) \in \mathbf{I},$$

find $\displaystyle\int_{\mathbf{I}} f$ and $\displaystyle\int_{\mathbf{I}} g$.

Question 2

Complete the details in the proof of Theorem 6.25.

Question 3

Complete the details in the proof of Theorem 6.26.

6.3 Jordan Measurable Sets and Riemann Integrable Functions

In this section, we will give some sufficient conditions for a bounded function $f : \mathfrak{D} \to \mathbb{R}$ to be Riemann integrable. We start with the following theorem.

Theorem 6.27

Let $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$, and let $f : \mathbf{I} \to \mathbb{R}$ be a continuous function defined on $\mathbf{I}$. Then $f : \mathbf{I} \to \mathbb{R}$ is Riemann integrable.

Proof

Since $f : \mathbf{I} \to \mathbb{R}$ is continuous and $\mathbf{I}$ is compact, $f : \mathbf{I} \to \mathbb{R}$ is uniformly continuous. Given $\varepsilon > 0$, there exists $\delta > 0$ such that if $\mathbf{u}$ and $\mathbf{v}$ are points in $\mathbf{I}$ and $\|\mathbf{u} - \mathbf{v}\| < \delta$, then

$$|f(\mathbf{u}) - f(\mathbf{v})| < \frac{\varepsilon}{\text{vol}\,(\mathbf{I})}.$$

Let $\mathbf{P}$ be any partition of $\mathbf{I}$ with $|\mathbf{P}| < \delta$. A rectangle $\mathbf{J}$ in $\mathcal{J}_\mathbf{P}$ is a compact set. Since $f : \mathbf{J} \to \mathbb{R}$ is continuous, the extreme value theorem says that there exist points $\mathbf{u}_\mathbf{J}$ and $\mathbf{v}_\mathbf{J}$ in $\mathbf{J}$ such that

$$f(\mathbf{u}_\mathbf{J}) \leq f(\mathbf{x}) \leq f(\mathbf{v}_\mathbf{J}) \qquad \text{for all } \mathbf{x} \in \mathbf{J}.$$

Therefore,

$$m_\mathbf{J} = \inf_{\mathbf{x} \in \mathbf{J}} f(\mathbf{x}) = f(\mathbf{u}_\mathbf{J}) \quad \text{and} \quad M_\mathbf{J} = \sup_{\mathbf{x} \in \mathbf{J}} f(\mathbf{x}) = f(\mathbf{v}_\mathbf{J}).$$

Since $|\mathbf{P}| < \delta$,

$$\|\mathbf{u}_\mathbf{J} - \mathbf{v}_\mathbf{J}\| \leq \text{diam}\,\mathbf{J} \leq |\mathbf{P}| < \delta.$$

Therefore,

$$M_\mathbf{J} - m_\mathbf{J} = f(\mathbf{v}_\mathbf{J}) - f(\mathbf{u}_\mathbf{J}) < \frac{\varepsilon}{\text{vol}\,(\mathbf{I})}.$$

This implies that

$$U(f, \mathbf{P}) - L(f, \mathbf{P}) = \sum_{\mathbf{J} \in \mathcal{J}_\mathbf{P}} (M_\mathbf{J} - m_\mathbf{J})\ \text{vol}\,(\mathbf{J}) < \frac{\varepsilon}{\text{vol}\,(\mathbf{I})} \sum_{\mathbf{J} \in \mathcal{J}_\mathbf{P}} \text{vol}\,(\mathbf{J}) = \varepsilon.$$

Hence, $f : \mathbf{I} \to \mathbb{R}$ is Riemann integrable.

Example 6.19

Let $f : [0,1] \times [0,1] \to \mathbb{R}$ be the function defined as

$$f(x,y) = \sin(xy).$$

This is a composition of the sine function and a polynomial, both of which are continuous functions. Hence, $f : [0,1] \times [0,1] \to \mathbb{R}$ is a continuous function. Therefore, $f : [0,1] \times [0,1] \to \mathbb{R}$ is Riemann integrable.

In Section 6.1, we have seen that a constant function $f : \mathbf{I} \to \mathbb{R}$, $f(\mathbf{x}) = c$ defined on $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$ is Riemann integrable and its integral is

$$\int_{\mathbf{I}} f = c \,\text{vol}\,(\mathbf{I}).$$

Since constant functions are the simplest bounded functions, it is natural to ask whether a constant function $f : \mathfrak{D} \to \mathbb{R}$, $f(\mathbf{x}) = c$ on a bounded set $\mathfrak{D}$ is always Riemann integrable. By linearity, it is sufficient to consider the case when $c = 1$. When $\mathfrak{D}$ is a closed rectangle of the form $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$, the answer is affirmative and we have

$$\int_{\mathbf{I}} d\mathbf{x} = \text{vol}\,(\mathbf{I})$$

To consider a general set $\mathfrak{D}$, let us first define the characteristic function of a set.

Definition 6.13 Characteristic Functions

Let A be a subset of $\mathbb{R}^n$. The characteristic function or indicator function of the set A is the function $\chi_A : \mathbb{R}^n \to \mathbb{R}$ defined as

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

Example 6.20

Let $A = \{(x, y) \mid x > 0\}$. Notice that the function $\chi_A : \mathbb{R}^2 \to \mathbb{R}$ is

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

It is continuous at (x, y) if and only if $x > 0$ or $x < 0$.

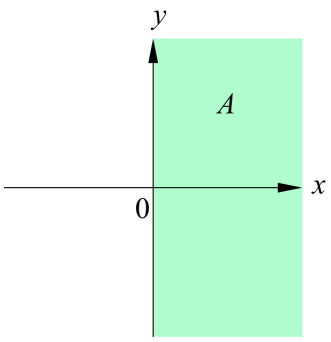


Figure 6.12: The set $A = \{(x, y) \mid x > 0\}$.

Interior, Exterior and Boundary of a Set

In Chapter 1, we have seen that if A is a subset of $\mathbb{R}^n$, then $\mathbb{R}^n$ is a disjoint union of int A, ext A and ∂A.
If $\mathbf{x}_0$ is a point in $\mathbb{R}^n$, $\mathbf{x}_0 \in \text{int}\, A$ if and only if there is an $r > 0$ such that $B(\mathbf{x}_0, r) \subset A$; $\mathbf{x}_0 \in \text{ext}\, A$ if and only if there is an $r > 0$ such that $B(\mathbf{x}_0, r) \subset \mathbb{R}^n \setminus A$; and $\mathbf{x}_0 \in \partial A$ if for every $r > 0$, $B(\mathbf{x}_0, r)$ contains a point in A and a point not in A.

Theorem 6.28

Let A be a subset of $\mathbb{R}^n$, and let $\chi_A : \mathbb{R}^n \to \mathbb{R}$ be the characteristic function of A. Then the set of discontinuities of the function χ_A is the set ∂A.

Proof

Since $\mathbb{R}^n$ is a disjoint union of int A, ext A and ∂A, we will show that χ_A is continuous on int A and ext A, and discontinuous at every point in ∂A.
The sets int A and ext A are open sets, and f is equal to 1 on int A and 0 on ext A. For every $\mathbf{x}_0$ in int A, there is an $r > 0$ such that $B(\mathbf{x}_0, r) \subset A$. Therefore, for any $\varepsilon > 0$, if $\mathbf{x}$ is such that $\|\mathbf{x} - \mathbf{x}_0\| < r$, then

$$|f(\mathbf{x}) - f(\mathbf{x}_0)| = 0 < \varepsilon.$$

This shows that f is continuous at $\mathbf{x}_0$. Similarly, if $\mathbf{x}_0$ is in ext A, there is an $r > 0$ such that $B(\mathbf{x}_0, r) \subset \mathbb{R}^n \setminus A$. The same reasoning shows that f is continuous at $\mathbf{x}_0$.
Now consider a point $\mathbf{x}_0$ that is in ∂A. For any $k \in \mathbb{Z}^+$, there is a point $\mathbf{u}_k \in A$ and a point $\mathbf{v}_k \notin A$ such that $\mathbf{u}_k$ and $\mathbf{v}_k$ are in the neighbourhood $B(\mathbf{x}_0, 1/k)$ of $\mathbf{x}_0$. The two sequences $\{\mathbf{u}_k\}$ and $\{\mathbf{v}_k\}$ both converge to $\mathbf{x}_0$, but the sequence $\{f(\mathbf{u}_k)\}$ converges to 1, the sequence $\{f(\mathbf{v}_k)\}$ converges to 0. This shows that f is not continuous at $\mathbf{x}_0$.

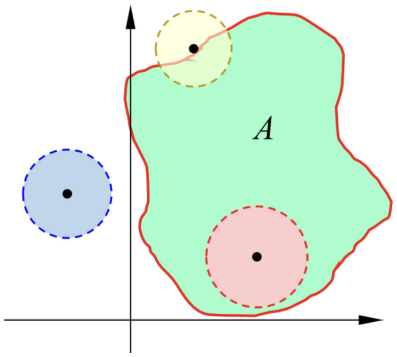


Figure 6.13: The characteristic function of a set A is not continuous at $\mathbf{x}_0$ if and only if $\mathbf{x}_0 \in \partial A$.

By definition, restricted to the set A, $\chi_A : A \to \mathbb{R}$ is the constant function $\chi_A(A) = 1$. Now we define Jordan measurable sets and its volume.

Definition 6.14 Jordan Measurable Sets and Volume

Let $\mathfrak{D}$ be a bounded subset of $\mathbb{R}^n$. We say that $\mathfrak{D}$ is Jordan measurable if the constant function $\chi_D : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable. In this case, we define the volume of $\mathfrak{D}$ as

$$\text{vol}\,(\mathfrak{D}) = \int_{\mathfrak{D}} \chi_{\mathfrak{D}} = \int_{\mathfrak{D}} d\mathbf{x}.$$

Example 6.21

The closed rectangle $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$ is Jordan measurable, and its volume is

$$\text{vol}\,(\mathbf{I}) = \int_{\mathbf{I}} d\mathbf{x} = \prod_{i=1}^{n} (b_i - a_i),$$

as what we have defined earlier.

Example 6.22

Example 6.14 says that the set

$$\mathfrak{D} = \{(x, y) \,|\, 0 \leq y \leq x \leq 1\}$$

is Jordan measurable and $\text{vol}\,(\mathfrak{D}) = \dfrac{1}{2}$.

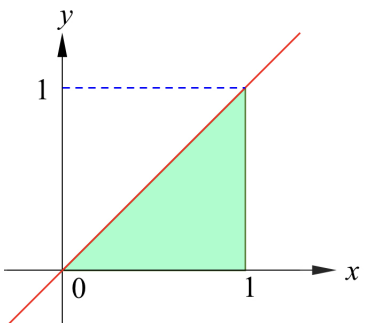


Figure 6.14: The set $\mathfrak{D} = \{(x, y) \,|\, 0 \leq y \leq x \leq 1\}$ is Jordan measurable.

One might think that all bounded subsets of $\mathbb{R}^n$ has volumes. This is not true. An example is given below.

Example 6.23

Let $\mathbf{I} = [0,1]^n$ and let

$$\mathfrak{D} = \{\mathbf{x} \in \mathbf{I} \,|\, \mathbf{x} \in \mathbb{Q}^n\}\,.$$

Notice that $\mathfrak{D}$ is a subset of the rectangle $\mathbf{I}$, and the zero extension of $\chi_{\mathfrak{D}} : \mathfrak{D} \to \mathbb{R}$ is the function $\chi_{\mathfrak{D}} : \mathbf{I} \to \mathbb{R}$,

$$\chi_{\mathfrak{D}}(\mathbf{x}) = \begin{cases} 1, & \text{if all components of } \mathbf{x} \text{ are rational,} \\ 0, & \text{otherwise,} \end{cases}$$

which is the Dirichlet's function. We have seen in Example 6.13 that the function $\chi_{\mathfrak{D}} : \mathbf{I} \to \mathbb{R}$ is not Riemann integrable. Hence, $\chi_{\mathfrak{D}} : \mathfrak{D} \to \mathbb{R}$ is not Riemann integrable. This means the set $\mathfrak{D}$ is not Jordan measurable and so it does not have a volume.

This example also shows that if B is a subset of A, and the function $f : A \to \mathbb{R}$ is Riemann integrable, the function $f : B \to \mathbb{R}$ is not necessary Riemann integrable.

The next example says that the boundary of a rectangle has volume 0.

Example 6.24

Let $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$, and let $\mathfrak{D} = \partial\mathbf{I}$. Notice that $\mathfrak{D}$ is contained in $\mathbf{I}$. The zero extension of $\chi_{\mathfrak{D}} : \mathfrak{D} \to \mathbb{R}$ is the function $\chi_{\mathfrak{D}} : \mathbf{I} \to \mathbb{R}$ which vanishes on the interior of $\mathbf{I}$. By Theorem 6.19, $\chi_{\mathfrak{D}} : \mathbf{I} \to \mathbb{R}$ is Riemann integrable and $\int_{\mathbf{I}} \chi_{\mathfrak{D}} = 0$. Therefore, $\mathfrak{D} = \partial\mathbf{I}$ has zero volume.

Remark 6.2 Darboux Sums for a Characteristic Function

Given a bounded set $\mathfrak{D}$ that is contained in the rectangle $\mathbf{I}$, if $\mathbf{P}$ is a partition of $\mathbf{I}$, $L(\chi_{\mathfrak{D}}, \mathbf{P})$ is the sum of the volumes of the rectangles in $\mathbf{P}$ that is contained in $\mathfrak{D}$; while $U(\chi_{\mathfrak{D}}, \mathbf{P})$ is the sum of the volumes of the rectangles in $\mathbf{P}$ that intersect $\mathfrak{D}$. See Figure 6.15.

Thus, for $\mathfrak{D}$ to have volume, the two numbers $L(\chi_{\mathfrak{D}}, \mathbf{P})$ and $U(\chi_{\mathfrak{D}}, \mathbf{P})$ should get closer and closer when the partitions $\mathbf{P}$ gets finer.

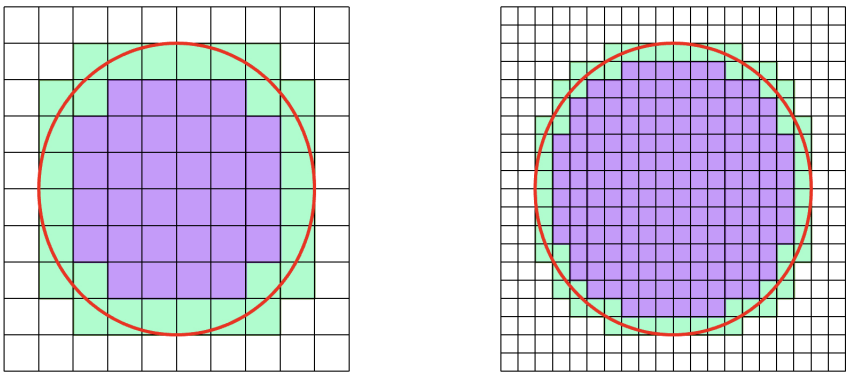


Figure 6.15: The geometric quantities represented by $L(\chi_{\mathfrak{D}}, \mathbf{P})$ and $U(\chi_{\mathfrak{D}}, \mathbf{P})$ when $\mathfrak{D}$ is the region bounded inside the circle.

Our goal is to give characterization of sets that are Jordan measurable. We will consider those that have zero volumes first. The following is a useful lemma.

Lemma 6.29

Let $\mathbf{I}$ be a closed rectangle in $\mathbb{R}^n$ that contains the closed rectangles $\mathbf{I}_1, \ldots, \mathbf{I}_k$. There is a partition $\mathbf{P}$ of $\mathbf{I}$ such that if $\mathbf{J}$ is a rectangle in the partition $\mathbf{P}$, then $\mathbf{J}$ is either contained in an $\mathbf{I}_j$ for some $1 \leq j \leq k$, or $\mathbf{J}$ is disjoint from the interiors of $\mathbf{I}_j$ for all $1 \leq j \leq k$.

Sketch of Proof

We construct the partition $\mathbf{P} = (P_1, \ldots, P_n)$ in the following way. For each $1 \leq i \leq n$, the partition points in P_i is the set of end points of the i^{th}-edge of $\mathbf{I}$, $\mathbf{I}_1$, …, $\mathbf{I}_k$. One can check that this partition satisfies the requirement. See Figure 6.16 for an illustration.

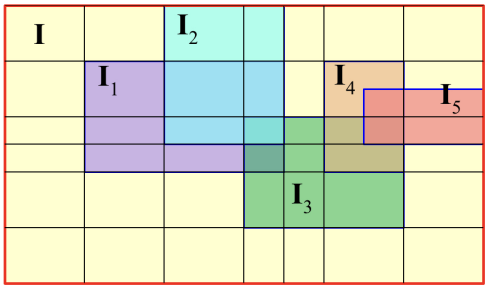


Figure 6.16: A partition of the rectangle $\mathbf{I}$ that satisfies the conditions in Lemma 6.29.

Let us introduce the definition of a cube.

Definition 6.15 Cubes

A rectangle of the form $\prod_{i=1}^{n}[a_i, b_i]$ such that

$$b_1 - a_1 = b_2 - a_2 = \cdots = b_n - a_n = \ell = 2r$$

is called a (closed) cube with side length $\ell = 2r$. The center of the cube is

$$\mathbf{c} = (c_1, c_2, \ldots, c_n) = \left(\frac{a_1 + b_1}{2}, \frac{a_2 + b_2}{2}, \ldots, \frac{a_n + b_n}{2}\right).$$

We will denote such a cube by $Q_{\mathbf{c},r}$.

There are also cubes whose edges are not parallel to the coordinate axes. In this chapter, when we say a cube, we always mean a cube defined above.

Now we can give a characterization of sets with zero volume.

Theorem 6.30

Let $\mathfrak{D}$ be a bounded subset of $\mathbb{R}^n$. The following are equivalent.

(a) The set $\mathfrak{D}$ is Jordan measurable and it has zero volume.

(b) For any $\varepsilon > 0$, there are finitely many closed cubes $Q_1, \ldots, Q_k$ such that
$$\mathfrak{D} \subset \bigcup_{j=1}^{k} Q_j \quad \text{and} \quad \sum_{j=1}^{k} \operatorname{vol}(Q_j) < \varepsilon.$$

(c) For any $\varepsilon > 0$, there are finitely many closed rectangles $\mathbf{I}_1, \ldots, \mathbf{I}_k$ such that
$$\mathfrak{D} \subset \bigcup_{j=1}^{k} \mathbf{I}_j \quad \text{and} \quad \sum_{j=1}^{k} \operatorname{vol}(\mathbf{I}_j) < \varepsilon.$$

Proof

First assume that $\mathfrak{D}$ is a Jordan measurable set with zero volume. There is a positive number R such that the closed cube $Q_{\mathbf{0},R} = [-R, R]^n$ contains the set $\mathfrak{D}$. Let $\mathbf{I} = Q_{\mathbf{0},R}$. Then the function $\chi_{\mathfrak{D}} : \mathbf{I} \to \mathbb{R}$ is Riemann integrable and $\int_{\mathbf{I}} \chi_{\mathfrak{D}} = 0$. Given $m \in \mathbb{Z}^+$, let $\mathbf{P}_m$ be the uniformly regular partition of $\mathbf{I}$ into m^n rectangles. Notice that each rectangle in the partition $\mathbf{P}_m$ is a cube. Since $\lim_{m\to\infty} |\mathbf{P}_m| = 0$, we have

$$\lim_{m\to\infty} U(\chi_{\mathfrak{D}}, \mathbf{P}_m) = \int_{\mathbf{I}} \chi_{\mathfrak{D}} = 0.$$

Given $\varepsilon > 0$, there is a positive integer M such that for all $m \geq M$,

$$U(\chi_{\mathfrak{D}}, \mathbf{P}_m) < \varepsilon.$$

Consider the partition $\mathbf{P}_M$. Notice that for $\mathbf{J} \in \mathcal{J}_{\mathbf{P}_M}$,

$$M_{\mathbf{J}}(\chi_{\mathfrak{D}}) = \begin{cases} 1, & \text{if } \mathbf{J} \cap \mathfrak{D} \neq \emptyset, \\ 0, & \text{if } \mathbf{J} \cap \mathfrak{D} = \emptyset. \end{cases}$$

Let

$$\mathscr{A} = \{\mathbf{J} \in \mathcal{J}_{\mathbf{P}_M} \,|\, \mathbf{J} \cap \mathfrak{D} \neq \emptyset\}\,.$$

Then

$$U(\chi_{\mathfrak{D}}, \mathbf{P}_M) = \sum_{\mathbf{J}\in\mathscr{A}} \text{vol}\,(\mathbf{J}).$$

$\mathscr{A}$ is a finite collection of cubess. Hence, we can named the cubes in $\mathscr{A}$ as $Q_1, \ldots, Q_k$. By construction,

$$\mathfrak{D} \subset \bigcup_{j=1}^{k} Q_j \quad \text{and} \quad \sum_{j=1}^{k} \text{vol}\,(Q_j) < \varepsilon.$$

This proves that (a) implies (b).
(b) implies (c) is obvious since a cube is a rectangle.
Now assume that (c) holds. Given $\varepsilon > 0$, (c) says that there are closed rectangles $\mathbf{I}_1, \ldots, \mathbf{I}_k$ such that

$$\mathfrak{D} \subset \bigcup_{j=1}^{k} \mathbf{I}_j \quad \text{and} \quad \sum_{j=1}^{k} \text{vol}\,(\mathbf{I}_j) < \frac{\varepsilon}{2}.$$

By Lemma 6.18, for each $1 \leq j \leq k$, there is a closed rectangle $\check{\mathbf{I}}_j$ such that $\mathbf{I}_j \subset \text{int}\,\check{\mathbf{I}}_j$ and

$$\text{vol}\,(\check{\mathbf{I}}_j) - \text{vol}\,(\mathbf{I}_j) < \frac{\varepsilon}{2k}.$$

It follows that

$$\mathfrak{D} \subset \bigcup_{j=1}^{k} \text{int}\,\check{\mathbf{I}}_j \quad \text{and} \quad \sum_{j=1}^{k} \text{vol}\,(\check{\mathbf{I}}_j) < \varepsilon.$$

Let $\mathbf{I}$ be a closed rectangle whose interior contains the bounded set $\bigcup_{j=1}^{k} \check{\mathbf{I}}_j$.

By Lemma 6.29, there is a partition $\mathbf{P}$ of $\mathbf{I}$ such that each rectangle $\mathbf{J}$ in the partition $\mathbf{P}$ is either contained in an $\check{\mathbf{I}}_j$ for some $1 \le j \le k$, or is disjoint from the interiors of $\check{\mathbf{I}}_j$ for all $1 \le j \le k$. Let

$$\mathscr{B} = \left\{ \mathbf{J} \in \mathcal{J}_{\mathbf{P}} \,|\, \mathbf{J} \subset \check{\mathbf{I}}_j \text{ for some } 1 \le j \le k \right\}.$$

If $\mathbf{J} \notin \mathscr{B}$, then $\mathbf{J} \cap \operatorname{int} \check{\mathbf{I}}_j = \emptyset$ for all $1 \le j \le k$. Therefore, $\mathbf{J} \cap \mathfrak{D} = \emptyset$. For these $\mathbf{J}$, $M_{\mathbf{J}}(\chi_{\mathfrak{D}}) = m_{\mathbf{J}}(\chi_{\mathfrak{D}}) = 0$. If $\mathbf{J}$ is in $\mathscr{B}$, we use the simple estimate $M_{\mathbf{J}} \le 1$. Thus,

$$U(\chi_{\mathfrak{D}}, \mathbf{P}) = \sum_{\mathbf{J} \in \mathscr{B}} M_{\mathbf{J}} \operatorname{vol}(\mathbf{J}) \le \sum_{\mathbf{J} \in \mathscr{B}} \operatorname{vol}(\mathbf{J}) \le \sum_{j=1}^{k} \operatorname{vol}(\check{\mathbf{I}}_j) < \varepsilon.$$

Since $L(\chi_{\mathfrak{D}}, \mathbf{P}) \ge 0$, we find that

$$U(\chi_{\mathfrak{D}}, \mathbf{P}) - L(\chi_{\mathfrak{D}}, \mathbf{P}) < \varepsilon.$$

This proves that $\chi_{\mathfrak{D}} : \mathbf{I} \to \mathbb{R}$ is Riemann integrable. Since we have shown that there exists a partition $\mathbf{P}$ such that $U(\chi_{\mathfrak{D}}, \mathbf{P}) < \varepsilon$, we have

$$\operatorname{vol}(\mathfrak{D}) = \int_{\mathbf{I}} \chi_{\mathfrak{D}} \le U(\chi_{\mathfrak{D}}, \mathbf{P}) < \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, we find that $\operatorname{vol}(\mathfrak{D}) = 0$. This completes the proof of (c) implies (a).

Motivated by Theorem 6.30, we make the following definition.

Definition 6.16 Jordan Content Zero

Let $\mathfrak{D}$ be a bounded subset of $\mathbb{R}^n$. We say that $\mathfrak{D}$ has Jordan content zero provided that for any $\varepsilon > 0$, there are finitely many closed rectangles $\mathbf{I}_1, \ldots, \mathbf{I}_k$ such that

$$\mathfrak{D} \subset \bigcup_{j=1}^{k} \mathbf{I}_j \quad \text{and} \quad \sum_{j=1}^{k} \operatorname{vol}(\mathbf{I}_j) < \varepsilon.$$

Sets that have Jordan Content Zero

Let $\mathfrak{D}$ be a bounded subset of $\mathbb{R}^n$. Theorem 6.30 says that $\mathfrak{D}$ is Jordan measurable with volume zero if and only if it has Jordan content zero.

The characterization of sets with zero volume given in Theorem 6.30 facilitates the proofs of properties of such sets.

Theorem 6.31

Let $\mathfrak{D}_1$ and $\mathfrak{D}_2$ be bounded subsets of $\mathbb{R}^n$. If $\mathfrak{D}_1$ has Jordan content zero and $\mathfrak{D}_2 \subset \mathfrak{D}_1$, then $\mathfrak{D}_2$ also has Jordan content zero.

Proof

Given $\varepsilon > 0$, since $\mathfrak{D}_1$ has Jordan content zero, there are closed rectangles $\mathbf{I}_1, \ldots, \mathbf{I}_k$ such that

$$\mathfrak{D}_1 \subset \bigcup_{j=1}^{k} \mathbf{I}_j \quad \text{and} \quad \sum_{j=1}^{k} \operatorname{vol}(\mathbf{I}_j) < \varepsilon.$$

Since $\mathfrak{D}_2 \subset \mathfrak{D}_1$, we find that

$$\mathfrak{D}_2 \subset \bigcup_{j=1}^{k} \mathbf{I}_j \quad \text{and} \quad \sum_{j=1}^{k} \operatorname{vol}(\mathbf{I}_j) < \varepsilon.$$

Therefore, $\mathfrak{D}_2$ also has Jordan content zero.

Example 6.25

Let $\mathfrak{D}$ be the subset of $\mathbb{R}^3$ given by

$$\mathfrak{D} = \{(x, y, 2) \mid -2 \leq x \leq 3, -5 \leq y \leq 7\}.$$

Show that $\mathfrak{D}$ is a Jordan measurable set with zero volume.

Solution

Let $\mathbf{I} = [-2,3] \times [-5,7] \times [2,3]$. Then $\mathbf{I}$ is a closed rectangle in $\mathbb{R}^3$. Example 6.24 says that $\partial\mathbf{I}$ has Jordan content zero. Since $\mathfrak{D} \subset \partial\mathbf{I}$, Theorem 6.31 says that $\mathfrak{D}$ has Jordan content zero. Hence, $\mathfrak{D}$ is a Jordan measurable set with zero volume.

The next theorem concerns unions and intersections of sets of Jordan content zero.

Theorem 6.32

(a) If $\mathscr{A} = \{\mathfrak{D}_\alpha \,|\, \alpha \in J\}$ is a collection of sets that have Jordan content zero, then their intersection $\mathcal{U} = \bigcap_{\alpha\in J} \mathfrak{D}_\alpha$ also has Jordan content zero.

(b) If $\mathfrak{D}_1, \ldots, \mathfrak{D}_m$ are finitely many sets that have Jordan content zero, then their union $\mathfrak{D} = \bigcup_{j=1}^{m} \mathfrak{D}_j$ is also a set that has Jordan content zero.

Proof

(a) is obvious since $\mathcal{U} \subset \mathfrak{D}_\alpha$ for any $\alpha \in J$.

(b) is basically a consequence of the fact that finite union of finite sets is finite. Given $\varepsilon > 0$, for each $1 \le j \le m$, since $\mathfrak{D}_j$ has Jordan content zero, there is a finite collection $\mathscr{B}_j = \{\mathbf{I}_{\beta_j} \,|\, \beta_j \in J_j\}$ of closed rectangles such that

$$\mathfrak{D}_j \subset \bigcup_{\beta_j\in J_j} \mathbf{I}_{\beta_j}, \qquad \sum_{\beta_j\in J_j} \text{vol}\left(\mathbf{I}_{\beta_j}\right) < \frac{\varepsilon}{m}.$$

Let

$$\mathscr{B} = \bigcup_{j=1}^{m} \mathscr{B}_j.$$

Since each $\mathscr{B}_j$, $1 \le j \le m$ is finite, $\mathscr{B}$ is also a finite collection of closed rectangles. Moreover,

$$\mathfrak{D} = \bigcup_{j=1}^{m} \mathfrak{D}_j \subset \bigcup_{j=1}^{m} \bigcup_{\beta_j\in J_j} \mathbf{I}_{\beta_j} = \bigcup_{\mathbf{I}_\beta\in\mathscr{B}} \mathbf{I}_\beta,$$

and

$$\sum_{\mathbf{I}_\beta \in \mathscr{B}} \text{vol}\,(\mathbf{I}_\beta) \leq \sum_{j=1}^{m} \sum_{\beta_j \in J_j} \text{vol}\,\left(\mathbf{I}_{\beta_j}\right) < \varepsilon.$$

This shows that $\mathfrak{D}$ has Jordan content zero.

Example 6.26

It is obvious that a one-point subset of $\mathbb{R}^n$ has Jordan content zero. It follows that any finite subset of $\mathbb{R}^n$ has Jordan content zero.

Now we want to consider general Jordan measurable sets. We first prove the following two theorems, giving more examples of Riemann integrable functions. The first one is a special case of the second one, but we need to prove it first to prove the second theorem.

Theorem 6.33

Let $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$, and let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function defined on $\mathbf{I}$. If $f : \mathbf{I} \to \mathbb{R}$ is continuous on the interior of $\mathbf{I}$, then $f : \mathbf{I} \to \mathbb{R}$ is Riemann integrable.

Proof

We will show that for any $\varepsilon > 0$, there is a partition $\mathbf{P}$ of $\mathbf{I}$ such that $U(f, \mathbf{P}) - L(f, \mathbf{P}) < \varepsilon$.

Since $f : \mathbf{I} \to \mathbb{R}$ is a bounded function, there is a positive number M such that

$$|f(\mathbf{x})| \leq M \qquad \text{for all } \mathbf{x} \in \mathbf{I}.$$

By Lemma 6.18, there is a closed rectangle $\check{\mathbf{I}} = \prod_{i=1}^{n} [u_i, v_i]$ contained in the interior of $\mathbf{I}$, such that

$$\text{vol}\,(\mathbf{I}) - \text{vol}\,(\check{\mathbf{I}}) < \frac{\varepsilon}{4M}.$$

Let $\mathbf{P}_0 = (P_1, \ldots, P_n)$ be the partition of $\mathbf{I}$ given by $P_i = \{a_i, u_i, v_i, b_i\}$ for $1 \leq i \leq n$. Then $\check{\mathbf{I}}$ is a rectangle in the partition $\mathbf{P}_0$.
Since $f : \check{\mathbf{I}} \to \mathbb{R}$ is continuous, Theorem 6.27 implies that there is a partition $\mathbf{P}_1$ of $\check{\mathbf{I}}$ such that

$$U(f, \mathbf{P}_1) - L(f, \mathbf{P}_1) < \frac{\varepsilon}{2}.$$

Let $\mathbf{P}$ be the partition of $\mathbf{I}$ so that it contains all the partition points in $\mathbf{P}_0$ and $\mathbf{P}_1$. Then $\mathbf{P}$ is a refinement of $\mathbf{P}_0$ and the partition that $\mathbf{P}$ induces on $\check{\mathbf{I}}$ is $\mathbf{P}(\check{\mathbf{I}}) = \mathbf{P}_1$. By Proposition 6.3,

$$\begin{aligned} U(f, \mathbf{P}) - L(f, \mathbf{P}) &= \sum_{\mathbf{J} \in \mathcal{J}_{\mathbf{P}_0}} \left(U(f, \mathbf{P}(\mathbf{J})) - L(f, \mathbf{P}(\mathbf{J}))\right) \\ &= U(f, \mathbf{P}_1) - L(f, \mathbf{P}_1) + \sum_{\mathbf{J} \in \mathcal{J}_{\mathbf{P}_0} \setminus \{\check{\mathbf{I}}\}} \left(U(f, \mathbf{P}(\mathbf{J})) - L(f, \mathbf{P}(\mathbf{J}))\right). \end{aligned}$$

For each $\mathbf{J}$ in $\mathcal{J}_{\mathbf{P}_0} \setminus \{\check{\mathbf{I}}\}$, we use the crude estimate

$$U(f, \mathbf{P}(\mathbf{J}) - L(f, \mathbf{P}(\mathbf{J})) \leq 2M \operatorname{vol}(\mathbf{J}).$$

These imply that

$$\begin{aligned} U(f, \mathbf{P}) - L(f, \mathbf{P}) &< \frac{\varepsilon}{2} + 2M \sum_{\mathbf{J} \in \mathcal{J}_{\mathbf{P}_0} \setminus \{\check{\mathbf{I}}\}} \operatorname{vol}(\mathbf{J}) \\ &= \frac{\varepsilon}{2} + 2M \left(\operatorname{vol}(\mathbf{I}) - \operatorname{vol}(\check{\mathbf{I}})\right) < \varepsilon. \end{aligned}$$

Set of Discontinuities of a Function

Given a function $f : A \to \mathbb{R}$ defined on the set A, the set of discontinuities of f is the set of all points $\mathbf{x}_0$ in A such that f is not continuous at $\mathbf{x}_0$.
If B is a subset of A, and $\mathbf{x}_0$ is a point of B, $f : A \to \mathbb{R}$ is continuous at $\mathbf{x}_0$ implies that $f : B \to \mathbb{R}$ is continuous at $\mathbf{x}_0$. Hence, the set of discontinuities of the function $f : B \to \mathbb{R}$ is a subset of the set of discontinuities of the function $f : A \to \mathbb{R}$.

Theorem 6.34

Let $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$. Given that $f : \mathbf{I} \to \mathbb{R}$ is a bounded function defined on $\mathbf{I}$, let $\mathcal{N}_f$ be the set of discontinuities of $f : \mathbf{I} \to \mathbb{R}$. If $\mathcal{N}_f$ is a set that has Jordan content zero, then $f : \mathbf{I} \to \mathbb{R}$ is Riemann integrable.

Proof

We will show that for any $\varepsilon > 0$, there is a partition $\mathbf{P}$ of $\mathbf{I}$ such that $U(f, \mathbf{P}) - L(f, \mathbf{P}) < \varepsilon$.

Since $f : \mathbf{I} \to \mathbb{R}$ is a bounded function, there is a positive number M such that

$$|f(\mathbf{x})| \leq M \qquad \text{for all } \mathbf{x} \in \mathbf{I}.$$

Since $\mathcal{N}_f$ is a set of Jordan content zero that is contained in $\mathbf{I}$, there are closed rectangles $\mathbf{I}_1, \ldots, \mathbf{I}_k$ such that

$$\mathcal{N}_f \subset \bigcup_{j=1}^{k} \mathbf{I}_j \quad \text{and} \quad \sum_{j=1}^{k} \text{vol}\,(\mathbf{I}_j) < \frac{\varepsilon}{4M}.$$

By Lemma 6.29, there is a partition $\mathbf{P}_0$ of $\mathbf{I}$ such that each rectangle $\mathbf{J}$ in the partition $\mathbf{P}_0$ is either contained in an $\mathbf{I}_j$ for some $1 \leq j \leq k$, or is disjoint from the interiors of $\mathbf{I}_j$ for all $1 \leq j \leq k$.

Let

$$\mathscr{A} = \{\mathbf{J} \in \mathcal{J}_{\mathbf{P}_0} \,|\, \mathbf{J} \subset \mathbf{I}_j \text{ for some } 1 \leq j \leq k\},$$

and

$$\mathscr{B} = \{\mathbf{J} \in \mathcal{J}_{\mathbf{P}_0} \,|\, \mathbf{J} \cap \text{int}\,(\mathbf{I}_j) = \emptyset \text{ for all } 1 \leq j \leq k\}.$$

Assume that $\mathscr{B}$ contains N rectangles. If $\mathbf{J} \in \mathscr{B}$, $f : \mathbf{J} \to \mathbb{R}$ is continuous on the interior of $\mathbf{J}$. By Theorem 6.33, $f : \mathbf{J} \to \mathbb{R}$ is Riemann integrable. Therefore, there is a partition $\mathbf{P}_\mathbf{J}$ of $\mathbf{J}$ such that

$$U(f, \mathbf{P}_\mathbf{J}) - L(f, \mathbf{P}_\mathbf{J}) < \frac{\varepsilon}{2N}.$$

The rest of the proof is similar to the proof of Theorem 6.33. Let $\mathbf{P}$ be the partition of $\mathbf{I}$ which contains all the partition points in $\mathbf{P}_0$ and $\mathbf{P}_\mathbf{J}$ for all $\mathbf{J} \in \mathscr{B}$. Then

$$\begin{aligned} U(f,\mathbf{P}) - L(f,\mathbf{P}) &= \sum_{\mathbf{J}\in\mathscr{J}_{\mathbf{P}_0}} \left(U(f,\mathbf{P}(\mathbf{J})) - L(f,\mathbf{P}(\mathbf{J}))\right) \\ &= \sum_{\mathbf{J}\in\mathscr{A}} \left(U(f,\mathbf{P}(\mathbf{J})) - L(f,\mathbf{P}(\mathbf{J}))\right) + \sum_{\mathbf{J}\in\mathscr{B}} \left(U(f,\mathbf{P}(\mathbf{J})) - L(f,\mathbf{P}(\mathbf{J}))\right). \end{aligned}$$

For each $\mathbf{J}$ in $\mathscr{A}$, we use the crude estimate

$$U(f,\mathbf{P}(\mathbf{J}) - L(f,\mathbf{P}(\mathbf{J})) \le 2M \operatorname{vol}(\mathbf{J}).$$

Using the fact that

$$\bigcup_{\mathbf{J}\in\mathscr{A}} \mathbf{J} \subset \bigcup_{j=1}^{k} \mathbf{I}_j,$$

we have

$$\sum_{\mathbf{J}\in\mathscr{A}} \left(U(f,\mathbf{P}(\mathbf{J})) - L(f,\mathbf{P}(\mathbf{J}))\right) \le 2M \sum_{j=1}^{k} \operatorname{vol}(\mathbf{I}_j) < \frac{\varepsilon}{2}.$$

For each $\mathbf{J} \in \mathscr{B}$, $\mathbf{P}(\mathbf{J})$ is a refinement of $\mathbf{P}_\mathbf{J}$, and thus

$$U(f,\mathbf{P}(\mathbf{J})) - L(f,\mathbf{P}(\mathbf{J})) < \frac{\varepsilon}{2N}.$$

This implies that

$$\sum_{\mathbf{J}\in\mathscr{B}} \left(U(f,\mathbf{P}(\mathbf{J})) - L(f,\mathbf{P}(\mathbf{J}))\right) < \frac{\varepsilon}{2}.$$

These give us $U(f,\mathbf{P}) - L(f,\mathbf{P}) < \varepsilon$, as desired.

Now we can prove the following characterization of Jordan measurable sets.

Theorem 6.35

Let $\mathfrak{D}$ be a bounded subset of $\mathbb{R}^n$. The following are equivalent.

(a) $\mathfrak{D}$ is a Jordan measurable set.

(b) The boundary of $\mathfrak{D}$ has Jordan content zero.

Proof

Let $\mathbf{I}$ be a closed rectangle that contains $\mathfrak{D}$. By definition, $\mathfrak{D}$ is Jordan measurable if and only if the function $\chi_{\mathfrak{D}} : \mathbf{I} \to \mathbb{R}$ is Riemann integrable. By Theorem 6.28, the set of discontinuities of the function $\chi_{\mathfrak{D}} : \mathbf{I} \to \mathbb{R}$ is the set $\partial\mathfrak{D}$. If the boundary of $\mathfrak{D}$ has Jordan content zero, Theorem 6.34 implies that $\chi_{\mathfrak{D}} : \mathbf{I} \to \mathbb{R}$ is Riemann integrable. This proves (b) implies (a). Conversely, if $\mathfrak{D}$ is Jordan measurable, given $\varepsilon > 0$, there is a partition $\mathbf{P}$ such that

$$U(\chi_{\mathfrak{D}}, \mathbf{P}) - L(\chi_{\mathfrak{D}}, \mathbf{P}) < \frac{\varepsilon}{2}.$$

For each $\mathbf{J}$ in $\mathcal{J}_{\mathbf{P}}$, there are only three possibilities for the pair $(m_{\mathbf{J}}, M_{\mathbf{J}})$. Namely, $(1,1)$, $(0,0)$ or $(0,1)$. Let

$$\begin{aligned}
\mathscr{A} &= \{\mathbf{J} \in \mathcal{J}_{\mathbf{P}} \,|\, m_{\mathbf{J}} = M_{\mathbf{J}} = 1\}, \\
\mathscr{B} &= \{\mathbf{J} \in \mathcal{J}_{\mathbf{P}} \,|\, m_{\mathbf{J}} = M_{\mathbf{J}} = 0\}, \\
\mathscr{C} &= \{\mathbf{J} \in \mathcal{J}_{\mathbf{P}} \,|\, m_{\mathbf{J}} = 0,\ M_{\mathbf{J}} = 1\}.
\end{aligned}$$

Then $\mathcal{J}_{\mathbf{P}} = \mathscr{A} \cup \mathscr{B} \cup \mathscr{C}$, and we have

$$U(\chi_{\mathfrak{D}}, \mathbf{P}) - L(\chi_{\mathfrak{D}}, \mathbf{P}) = \sum_{\mathbf{J} \in \mathcal{J}_{\mathbf{P}}} (M_{\mathbf{J}} - m_{\mathbf{J}}) \operatorname{vol}(\mathbf{J}) = \sum_{\mathbf{J} \in \mathscr{C}} \operatorname{vol}(\mathbf{J}).$$

This implies that

$$\sum_{\mathbf{J} \in \mathscr{C}} \operatorname{vol}(\mathbf{J}) < \frac{\varepsilon}{2}.$$

Notice that $\mathbf{J}$ is in $\mathscr{A}$ if and only if $f(\mathbf{x}) = 1$ for all $\mathbf{x} \in \mathbf{J}$, if and only if $\mathbf{J} \subset \mathfrak{D}$. This implies that

$$\operatorname{int} \mathbf{J} \subset \operatorname{int} \mathfrak{D} \qquad \text{for all } \mathbf{J} \in \mathscr{A}.$$

Similarly, $\mathbf{J}$ is in $\mathscr{B}$ if and only if $f(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathbf{J}$, if and only if $\mathbf{J} \subset \mathbb{R}^n \setminus \mathfrak{D}$. This implies that

$$\operatorname{int} \mathbf{J} \subset \operatorname{ext} \mathfrak{D} \qquad \text{for all } \mathbf{J} \in \mathscr{B}.$$

Let

$$\mathcal{S} = \bigcup_{\mathbf{J} \in \mathscr{A} \cup \mathscr{B}} \partial \mathbf{J}.$$

Since $\mathbb{R}^n$ is a disjoint union of $\operatorname{int} \mathfrak{D}$, $\operatorname{ext} \mathfrak{D}$ and $\partial \mathfrak{D}$, we must have

$$\partial \mathfrak{D} \subset \left(\bigcup_{\mathbf{J} \in \mathscr{C}} \mathbf{J} \right) \cup \mathcal{S}.$$

Since the boundary of a closed rectangle has Jordan content zero, and $\mathscr{A} \cup \mathscr{B}$ is a finite set, Theorem 6.32 implies that $\mathcal{S}$ has Jordan content zero. Hence, there is a finite collection of rectangles $\mathscr{D} = \{\mathbf{I}_j \,|\, 1 \leq j \leq k\}$ such that

$$\mathcal{S} \subset \bigcup_{j=1}^{k} \mathbf{I}_j \quad \text{and} \quad \sum_{j=1}^{k} \operatorname{vol}(\mathbf{I}_j) < \frac{\varepsilon}{2}.$$

Let $\mathscr{E} = \mathscr{C} \cup \mathscr{D}$. Then $\mathscr{E}$ is a finite collection of closed rectangles,

$$\partial \mathfrak{D} \subset \bigcup_{\mathbf{J} \in \mathscr{E}} \mathbf{J} \quad \text{and} \quad \sum_{\mathbf{J} \in \mathscr{E}} \operatorname{vol}(\mathbf{J}) < \varepsilon.$$

This shows that $\partial \mathfrak{D}$ has Jordan content zero.

Using Theorem 6.35, we can obtain more examples of Jordan measurable sets. First we prove the following.

Lemma 6.36

Let A and B be subsets of $\mathbb{R}^n$. Then

$$\partial(A \cup B) \subset \partial A \cup \partial B, \qquad \partial(A \cap B) \subset \partial A \cup \partial B.$$

Proof

If $\mathbf{x}_0$ be a point in $\partial(A \cup B)$, there is sequence of points $\{\mathbf{u}_k\}$ in $A \cup B$ that converges to $\mathbf{x}_0$. Each point in this sequence is either in A or in B. Therefore, there is a subsequence $\{\mathbf{u}_{k_j}\}$ that is in A or in B. There is also a sequence $\{\mathbf{v}_k\}$ in $\mathbb{R}^n \setminus (A \cup B)$ that converges to $\mathbf{x}_0$. This sequence is in both $\mathbb{R}^n \setminus A$ and in $\mathbb{R}^n \setminus B$. Therefore, $\mathbf{x}_0$ is in ∂A or in ∂B.
If $\mathbf{x}_0$ is a point in $\partial(A \cap B)$, there is sequence of points $\{\mathbf{u}_k\}$ in $\mathbb{R}^n \setminus (A \cap B)$ that converges to $\mathbf{x}_0$. Each point in this sequence is either in $\mathbb{R}^n \setminus A$ or in $\mathbb{R}^n \setminus B$. Therefore, there is a subsequence $\{\mathbf{u}_{k_j}\}$ that is in $\mathbb{R}^n \setminus A$ or in $\mathbb{R}^n \setminus B$. There is also a sequence $\{\mathbf{v}_k\}$ in $A \cap B$ that converges to $\mathbf{x}_0$. This sequence is in both A and B. Therefore, $\mathbf{x}_0$ is in ∂A or in ∂B.

One is tempted to think that $\partial(A \cap B) \subset \partial A \cap \partial B$. But this is not true, as shown in the following example.

Example 6.27

Let $A = [0,2] \times [0,2]$ and $B = [1,3] \times [1,3]$. We find that $A \cap B = [1,2] \times [1,2]$. As shown in Figure 6.17, $\partial A \cap \partial B$ is a set with 4 points, $\partial(A \cap B) \neq \partial A \cap \partial B$, but $\partial(A \cap B) \subset \partial A \cup \partial B$.

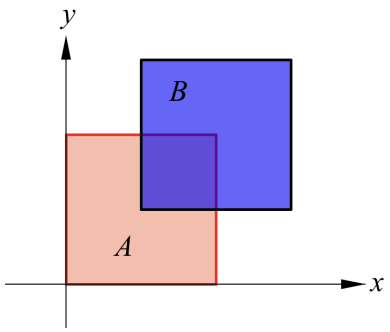


Figure 6.17: $\partial(A \cap B) \neq \partial A \cap \partial B$, but $\partial(A \cap B) \subset \partial A \cup \partial B$.

Using Lemma 6.36, we obtain the following.

Theorem 6.37

If $\mathfrak{D}_1, \mathfrak{D}_2, \ldots, \mathfrak{D}_m$ are Jordan measurable sets, then the set $\mathfrak{D}_1 \cap \mathfrak{D}_2 \cap \cdots \cap \mathfrak{D}_m$ and the set $\mathfrak{D}_1 \cup \mathfrak{D}_2 \cup \cdots \cup \mathfrak{D}_m$ are also Jordan measurable.

Proof

It suffices to prove the case where $m = 2$. The general case follows by induction.

If $\mathfrak{D}_1$ and $\mathfrak{D}_2$ are Jordan measurable, Theorem 6.35 says that $\partial\mathfrak{D}_1$ and $\partial\mathfrak{D}_2$ have Jordan content zero. Theorem 6.32 says that $\partial\mathfrak{D}_1 \cup \partial\mathfrak{D}_2$ has Jordan content zero. Lemma 6.36 and Theorem 6.31 imply that $\partial(\mathfrak{D}_1 \cap \mathfrak{D}_2)$ and $\partial(\mathfrak{D}_1 \cup \mathfrak{D}_2)$ have Jordan content zero. Theorem 6.35 again implies that $\mathfrak{D}_1 \cap \mathfrak{D}_2$ and $\mathfrak{D}_1 \cup \mathfrak{D}_2$ are Jordan measurable sets.

Observe that the concept of Jordan measurable sets and Riemann integrable functions are closely related. In a nutshell, a set $\mathfrak{D}$ is a Jordan measurable set if and only if all the constant functions $f : \mathfrak{D} \to \mathbb{R}$ are Riemann integrable. In fact, it is also if and only if all the continuous functions $f : \mathfrak{D} \to \mathbb{R}$ are Riemann integrable. We can even allow discontinuities on a set that has Jordan content zero. We will prove this after a few preparatory remarks and lemmas.

Remark 6.3

If $\mathfrak{D}$ is a bounded subset of $\mathbb{R}^n$ and $\mathfrak{D}$ is contained in the closed rectangle $\mathbf{I}$, then $\overline{\mathfrak{D}}$ is also contained in $\mathbf{I}$. This implies that $\partial\mathfrak{D}$ is also contained in $\mathbf{I}$.

The following example depicts the relation between the set of dicontinuities of a function $f : \mathfrak{D} \to \mathbb{R}$ and its zero extension.

Example 6.28

Let $\mathfrak{D} = \{(x, y) \,|\, 0 \leq y \leq x \leq 2\}$, and let $f : \mathfrak{D} \to \mathbb{R}$ be the function defined as

$$f(x, y) = \begin{cases} 1, & \text{if } (x, y) \in \mathfrak{D} \text{ and } 0 \leq x < 1, \\ 2, & \text{if } (x, y) \in \mathfrak{D} \text{ and } 1 \leq x \leq 2. \end{cases}$$

The set of discontinuities of the function $f : \mathfrak{D} \to \mathbb{R}$ is

$$\mathcal{N}_f = \{(1, y) \,|\, 0 \le y \le x\}.$$

The set $\mathfrak{D}$ is contained in the rectangle $\mathbf{I} = [0, 2] \times [0, 2]$. Let $\check{f} : \mathbf{I} \to \mathbb{R}$ be the zero extension of $f : \mathfrak{D} \to \mathbb{R}$. Then the set of discontinuities of $\check{f} : \mathbf{I} \to \mathbb{R}$ is

$$\mathcal{N}_{\check{f}} = \{(1, y) \,|\, 0 \le y \le 1\} \cup \{(x, x) \,|\, 0 \le x \le 2\},$$

which is a subset of $\mathcal{N}_f \cup \partial\mathfrak{D}$.

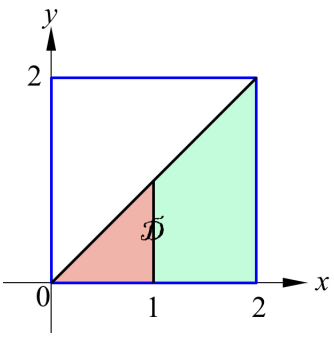


Figure 6.18: The set of discontinuities of the functions discussed in Example 6.28.

The following lemma gives the general case.

Lemma 6.38

Let $\mathfrak{D}$ be a bounded subset of $\mathbb{R}^n$. Given that $f : \mathfrak{D} \to \mathbb{R}$ is a bounded function defined on $\mathfrak{D}$, let $\check{f} : \mathbb{R}^n \to \mathbb{R}$ be its zero extension. Let

$$\mathcal{N}_f = \{\mathbf{x}_0 \in \mathfrak{D} \,|\, f : \mathfrak{D} \to \mathbb{R} \text{ is not continuous at } \mathbf{x}_0\},$$
$$\mathcal{N}_{\check{f}} = \{\mathbf{x}_0 \in \mathbb{R}^n \,|\, \check{f} : \mathbb{R}^n \to \mathbb{R} \text{ is not continuous at } \mathbf{x}_0\}.$$

Then

$$\mathcal{N}_{\check{f}} \subset \partial\mathfrak{D} \cup \mathcal{N}_f.$$

Proof

We will show that $\check{f} : \mathbb{R}^n \to \mathbb{R}$ is continuous at $\mathbf{x}_0$ if $\mathbf{x}_0$ is not in $\partial\mathfrak{D} \cup \mathcal{N}_f$. If $\mathbf{x}_0 \notin \partial\mathfrak{D} \cup \mathcal{N}_f$, there are two possibilities.

- $\mathbf{x}_0$ is in ext $\mathfrak{D}$.
- $\mathbf{x}_0$ is in int $\mathfrak{D}$ and $f : \mathfrak{D} \to \mathbb{R}$ is continuous at $\mathbf{x}_0$.

If $\mathbf{x}_0$ is in ext $\mathfrak{D}$, there is an $r > 0$ such that $B(\mathbf{x}_0, r) \subset \mathbb{R}^n \setminus \mathfrak{D}$. Since $\check{f}(\mathbf{x}) = 0$ for all $\mathbf{x} \in B(\mathbf{x}_0, r)$, $\check{f} : \mathbb{R}^n \to \mathbb{R}$ is continuous at $\mathbf{x}_0$.
If $\mathbf{x}_0$ is in int $\mathfrak{D}$ and $f : \mathfrak{D} \to \mathbb{R}$ is continuous at $\mathbf{x}_0$, we want to show that $\check{f} : \mathbb{R}^n \to \mathbb{R}$ is continuous at $\mathbf{x}_0$. Given that $\{\mathbf{x}_k\}$ is a sequence in $\mathbb{R}^n$ that converges to $\mathbf{x}_0$, there is a positive integer K such that $\mathbf{x}_k \in$ int $\mathfrak{D}$ for all $k \geq K$. This implies that $\check{f}(\mathbf{x}_k) = f(\mathbf{x}_k)$ for all $k \geq K$. Since $f : \mathfrak{D} \to \mathbb{R}$ is continuous at $\mathbf{x}_0$,

$$\lim_{k\to\infty} f(\mathbf{x}_{K+k}) = f(\mathbf{x}_0).$$

This implies that

$$\lim_{k\to\infty} \check{f}(\mathbf{x}_k) = \lim_{k\to\infty} \check{f}(\mathbf{x}_{K+k}) = \lim_{k\to\infty} f(\mathbf{x}_{K+k}) = f(\mathbf{x}_0) = \check{f}(\mathbf{x}_0).$$

Hence, $\check{f} : \mathbb{R}^n \to \mathbb{R}$ is continuous at $\mathbf{x}_0$. This completes the proof.

Now we can prove the main theorem.

Theorem 6.39

Let $\mathfrak{D}$ be a bounded subset of $\mathbb{R}^n$. Given that $f : \mathfrak{D} \to \mathbb{R}$ is a bounded function defined on $\mathfrak{D}$, let $\mathcal{N}_f$ be the set of discontinuities of $f : \mathfrak{D} \to \mathbb{R}$. If $\mathfrak{D}$ is a Jordan measurable set, and $\mathcal{N}_f$ is a set that has Jordan content zero, then $f : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable.

Proof

Let $\mathbf{I}$ be a closed rectangle in $\mathbb{R}^n$ that contains $\mathfrak{D}$, and let $\check{f} : \mathbf{I} \to \mathbb{R}$ be the zero extension of $f : \mathfrak{D} \to \mathbb{R}$ to $\mathbf{I}$. We want to show that $\check{f} : \mathbf{I} \to \mathbb{R}$ is Riemann integrable.

By Lemma 6.38, the set of discontinuities of $\check{f} : \mathbf{I} \to \mathbb{R}$ is contained in $\partial\mathfrak{D} \cup \mathcal{N}_f$. Since $\mathfrak{D}$ is Jordan measurable, Theorem 6.35 says that $\partial\mathfrak{D}$ has Jordan content zero. Theorem 6.32 and Theorem 6.31 then imply that the set $\mathcal{N}_{\check{f}}$ has Jordan content zero. By Theorem 6.34, $\check{f} : \mathbf{I} \to \mathbb{R}$ is Riemann integrable. This completes the proof.

In particular, we have the following.

Corollary 6.40

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$ that is Jordan measurable, and let $f : \mathfrak{D} \to \mathbb{R}$ be a bounded function defined on $\mathfrak{D}$. If $f : \mathfrak{D} \to \mathbb{R}$ is continuous, then it is Riemann integrable.

Let us emphasize the result in Corollary 6.40.

Riemann Integrability of Continuous Functions

Any continuous function defined on a Jordan measurable set is Riemann integrable.

Another interesting corollary of Theorem 6.39 is the following.

Corollary 6.41

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$ that has Jordan content zero. If $f : \mathfrak{D} \to \mathbb{R}$ is a bounded function defined on $\mathfrak{D}$, then $f : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable and

$$\int_{\mathfrak{D}} f \;=\; 0.$$

Proof

Since $\mathfrak{D}$ has Jordan content zero, and the set $\mathcal{N}_f$ of discontinuities of $f : \mathfrak{D} \to \mathbb{R}$ is a subset of $\mathfrak{D}$, $\mathcal{N}_f$ has Jordan content zero. Theorem 6.39 implies that $f : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable. Since $f : \mathfrak{D} \to \mathbb{R}$ is bounded, there is a positive number M such that $-M \leq f(\mathbf{x}) \leq M$ for all $\mathbf{x} \in \mathfrak{D}$.

By monotonicity theorem,

$$\int_{\mathfrak{D}} -M \, d\mathbf{x} \;\leq\; \int_{\mathfrak{D}} f(\mathbf{x}) \, d\mathbf{x} \;\leq\; \int_{\mathfrak{D}} M \, d\mathbf{x}.$$

For any constant c, linearity implies that $\displaystyle\int_{\mathfrak{D}} c \, d\mathbf{x} \;=\; c \operatorname{vol}(\mathfrak{D}) = 0$. This proves that

$$\int_{\mathfrak{D}} f \;=\; 0.$$

Let us highlight this important result.

Bounded Functions Defined on Sets that has Jordan Content Zero

Any bounded function defined on a set that has Jordan content zero is Riemann integrable with integral zero.

Corollary 6.41 also gives the following.

Lemma 6.42

Let $\mathfrak{D}$ be a bounded subset of $\mathbb{R}^n$, and let $f : \mathfrak{D} \to \mathbb{R}$ be a bounded function defined on $\mathfrak{D}$. If there is a subset $\mathcal{A}$ of $\mathfrak{D}$ with Jordan content zero such that $f(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathfrak{D} \setminus \mathcal{A}$, then $f : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable and $\displaystyle\int_{\mathfrak{D}} f = 0$.

Proof

Let $\mathbf{I}$ be a closed rectangle in $\mathbb{R}^n$ that contains $\mathfrak{D}$, and let $\check{f} : \mathbf{I} \to \mathbb{R}$ be the zero extension of the function $f : \mathfrak{D} \to \mathbb{R}$. Notice that it is also the zero extension of the function $f|_{\mathcal{A}} : \mathcal{A} \to \mathbb{R}$. By Corollary 6.41, $f|_{\mathcal{A}} : \mathcal{A} \to \mathbb{R}$ is Riemann integrable and $\displaystyle\int_{\mathcal{A}} f|_{\mathcal{A}} = 0$. Hence, $\check{f} : \mathbf{I} \to \mathbb{R}$ is Riemann integrable, and so is $f : \mathfrak{D} \to \mathbb{R}$. Moreover,

$$\int_{\mathfrak{D}} f = \int_{\mathbf{I}} \check{f} = \int_{\mathcal{A}} f|_{\mathcal{A}} = 0.$$

Using Lemma 6.42, we obtain the following important result, which says that

Riemann integrability is not affected by the definition of the function on a set that has Jordan content zero.

Theorem 6.43

Let $\mathfrak{D}$ be a bounded subset of $\mathbb{R}^n$, and let $f : \mathfrak{D} \to \mathbb{R}$ and $g : \mathfrak{D} \to \mathbb{R}$ be bounded functions defined on $\mathfrak{D}$. If $f : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable and there is a subset $\mathcal{A}$ of $\mathfrak{D}$ which has Jordan content zero such that

$$g(\mathbf{x}) = f(\mathbf{x}) \qquad \text{for all } \mathbf{x} \in \mathfrak{D} \setminus \mathcal{A},$$

then $g : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable and

$$\int_{\mathfrak{D}} g = \int_{\mathfrak{D}} f.$$

Sketch of Proof

Let $h : \mathfrak{D} \to \mathbb{R}$ be the function $h(\mathbf{x}) = f(\mathbf{x}) - g(\mathbf{x})$. Then $h(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathfrak{D} \setminus \mathcal{A}$. By Lemma 6.42, $h : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable and $\int_{\mathfrak{D}} h = 0$. The assertion follows from linearity.

Using Theorem 6.43, we can generalize additivity to arbitrary sets.

Theorem 6.44 Additivity

Given that $\mathfrak{D}_1$ and $\mathfrak{D}_2$ are bounded subsets of $\mathbb{R}^n$ such that $\mathfrak{D}_1 \cap \mathfrak{D}_2$ is a set that has Jordan content zero, let $\mathfrak{D} = \mathfrak{D}_1 \cup \mathfrak{D}_2$. Assume that $f : \mathfrak{D} \to \mathbb{R}$ is a bounded function defined on $\mathfrak{D}$. If the functions $f : \mathfrak{D}_1 \to \mathbb{R}$ and $f : \mathfrak{D}_2 \to \mathbb{R}$ are Riemann integrable, then the function $f : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable and

$$\int_{\mathfrak{D}} f = \int_{\mathfrak{D}_1 \cup \mathfrak{D}_2} f = \int_{\mathfrak{D}_1} f + \int_{\mathfrak{D}_2} f.$$

Proof

Let $\mathbf{I}$ be a closed rectangle in $\mathbb{R}^n$ that contains $\mathfrak{D}$, and let $\mathfrak{D}_0 = \mathfrak{D}_1 \cap \mathfrak{D}_2$. We are given that $\mathfrak{D}_0$ has Jordan content zero. Let $\check{f} : \mathbf{I} \to \mathbb{R}$, $\check{f}_1 : \mathbf{I} \to \mathbb{R}$, $\check{f}_2 : \mathbf{I} \to \mathbb{R}$ and $\check{f}_0 : \mathbf{I} \to \mathbb{R}$ be resepctively the zero extensions of $f : \mathfrak{D} \to \mathbb{R}$, $f : \mathfrak{D}_1 \to \mathbb{R}$, $f : \mathfrak{D}_2 \to \mathbb{R}$ and $f : \mathfrak{D}_0 \to \mathbb{R}$. It is easy to see that

$$\check{f}(\mathbf{x}) = \check{f}_1(\mathbf{x}) + \check{f}_2(\mathbf{x}) - \check{f}_0(\mathbf{x}) \qquad \text{for all } \mathbf{x} \in \mathbf{I}. \tag{6.3}$$

By Corollary 6.41, $f : \mathfrak{D}_0 \to \mathbb{R}$ is Riemann integrable and $\displaystyle\int_{\mathfrak{D}_0} f = 0$. Hence, $\check{f}_0 : \mathbf{I} \to \mathbb{R}$ is Riemann integrable.
Since $f : \mathfrak{D}_1 \to \mathbb{R}$ and $f : \mathfrak{D}_2 \to \mathbb{R}$ are Riemann integrable, $\check{f}_1 : \mathbf{I} \to \mathbb{R}$, $\check{f}_2 : \mathbf{I} \to \mathbb{R}$ are Riemann integrable. Linearly and (6.3) imply that $\check{f} : \mathbf{I} \to \mathbb{R}$ is Riemann integrable and

$$\int_{\mathfrak{D}} f = \int_{\mathfrak{D}_1} f + \int_{\mathfrak{D}_2} f - \int_{\mathfrak{D}_0} f = \int_{\mathfrak{D}_1} f + \int_{\mathfrak{D}_2} f.$$

By induction and Theorem 6.32, we obtain the following.

Theorem 6.45

Given that $\mathfrak{D}_1$, $\mathfrak{D}_2$, …, $\mathfrak{D}_m$ are bounded subsets of $\mathbb{R}^n$ such that for any pairs of (i, j) with $i \neq j$, $\mathfrak{D}_i \cap \mathfrak{D}_j$ has Jordan content zero, let $\mathfrak{D} = \mathfrak{D}_1 \cup \mathfrak{D}_2 \cup \cdots \cup \mathfrak{D}_m$. Assume that $f : \mathfrak{D} \to \mathbb{R}$ is a bounded function defined on $\mathfrak{D}$. If the functions $f : \mathfrak{D}_j \to \mathbb{R}$, $1 \leq j \leq m$, are Riemann integrable, then the function $f : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable and

$$\int_{\mathfrak{D}} f = \int_{\mathfrak{D}_1 \cup \mathfrak{D}_2 \cup \cdots \cup \mathfrak{D}_m} f = \int_{\mathfrak{D}_1} f + \int_{\mathfrak{D}_2} f + \cdots + \int_{\mathfrak{D}_m} f.$$

Remark 6.4

If a function $f : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable and $\mathfrak{D}_1$ is a subset of $\mathfrak{D}$, $f : \mathfrak{D}_1 \to \mathbb{R}$ is not necessarily Riemann integrable. For example, consider the constant function f on $\mathfrak{D} = [0, 1]^n$ which takes value 1. Its restriction to $\mathfrak{D}_1 = \mathfrak{D} \cap \mathbb{Q}^n$ is not Riemann integrable.
Thus for the general additivity theorem, we do not have if and only if.

Using the fact that a set $\mathfrak{D}$ is Jordan measurable if the function $\chi_{\mathfrak{D}} : \mathfrak{D} \to \mathbb{R}$ is measurable, Theorem 6.45 gives the following.

Corollary 6.46

Let $\mathfrak{D}_1, \mathfrak{D}_2, \ldots, \mathfrak{D}_m$ be Jordan measurable subsets of $\mathbb{R}^n$ such that for any pairs of (i, j) with $i \neq j$, $\mathfrak{D}_i \cap \mathfrak{D}_j$ has Jordan content zero. Then the set $\mathfrak{D} = \mathfrak{D}_1 \cup \mathfrak{D}_2 \cup \cdots \cup \mathfrak{D}_m$ is also Jordan measurable. Moreover,

$$\text{vol}\,(\mathfrak{D}) = \text{vol}\,(\mathfrak{D}_1) + \text{vol}\,(\mathfrak{D}_2) + \cdots + \text{vol}\,(\mathfrak{D}_m).$$

Let us return to explore more on Jordan measurable sets. So far we only know explicitly that a closed rectangle is Jordan measurable, and some examples of sets that have Jordan content zero. Since a bounded subset $\mathfrak{D}$ is Jordan measurable if and only if its boundary has Jordan content zero, we will first explore sets that have Jordan content zero. The following theorem will give us a lots of examples of sets that have Jordan content zero.

Theorem 6.47

Let $\mathfrak{D}$ be a Jordan measurable subset of $\mathbb{R}^n$, and let $f : \mathfrak{D} \to \mathbb{R}$ be a Riemann integrable function. Then the graph of f defined by

$$G_f = \left\{(\mathbf{x}, y) \in \mathbb{R}^{n+1} \,|\, \mathbf{x} \in \mathfrak{D}, y = f(\mathbf{x})\right\}$$

is a subset of $\mathbb{R}^{n+1}$ that has Jordan content zero.

Proof

Let $\mathbf{I}$ be a closed rectangle in $\mathbb{R}^n$ that contains $\mathfrak{D}$, and let $\check{f} : \mathbf{I} \to \mathbb{R}$ be the zero extension of $f : \mathfrak{D} \to \mathbb{R}$. Fixed $\varepsilon > 0$. Since $f : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable, there is a partition $\mathbf{P}$ of $\mathbf{I}$ such that

$$U(\check{f}, \mathbf{I}) - L(\check{f}, \mathbf{I}) < \frac{\varepsilon}{2}.$$

Let

$$\eta = \frac{\varepsilon}{4\text{vol}\,(\mathbf{I})}.$$

Then $\eta > 0$. Let

$$\mathscr{A} = \{\mathbf{J} \times [m_{\mathbf{J}} - \eta, M_{\mathbf{J}} + \eta] \,|\, \mathbf{J} \in \mathcal{J}_{\mathbf{P}}\}.$$

Then $\mathscr{A}$ is a finite collection of closed rectangles in $\mathbb{R}^{n+1}$. If $(\mathbf{x}, f(\mathbf{x}))$ is in G_f, there is a $\mathbf{J} \in \mathcal{J}_{\mathbf{P}}$ such that $\mathbf{x} \in \mathbf{J}$. Then $m_{\mathbf{J}} \leq f(\mathbf{x}) \leq M_{\mathbf{J}}$ implies that $(\mathbf{x}, f(\mathbf{x}))$ is in $\mathbf{J} \times [m_{\mathbf{J}} - \eta, M_{\mathbf{J}} + \eta]$. This proves that

$$G_f \subset \bigcup_{\mathbf{K} \in \mathscr{A}} \mathbf{K}.$$

Now,

$$\begin{aligned}\sum_{\mathbf{K} \in \mathscr{A}} \text{vol}\,(\mathbf{K}) &= \sum_{\mathbf{J} \in \mathcal{J}_{\mathbf{P}}} (M_{\mathbf{J}} - m_{\mathbf{J}} + 2\eta)\ \text{vol}\,(\mathbf{J}) \\ &= U(\check{f}, \mathbf{P}) - L(\check{f}, \mathbf{P}) + 2\eta \sum_{\mathbf{J} \in \mathcal{J}_{\mathbf{P}}} \text{vol}\,(\mathbf{J}) < \frac{\varepsilon}{2} + 2\eta\, \text{vol}\,(\mathbf{I}) < \varepsilon.\end{aligned}$$

This proves that G_f has Jordan content zero.

Specialize to continuous functions, we have the following.

Corollary 6.48

Let $\mathfrak{D}$ be a Jordan measurable set in $\mathbb{R}^n$, and let $f : \mathfrak{D} \to \mathbb{R}$ be a continuous function. Then the graph of f defined by

$$G_f = \left\{(\mathbf{x}, y) \in \mathbb{R}^{n+1} \,|\, \mathbf{x} \in \mathfrak{D}, y = f(\mathbf{x})\right\}$$

is a subset of $\mathbb{R}^{n+1}$ that has Jordan content zero.

Example 6.29

Any line segment L between two points (x_1, y_1) and (x_2, y_2) in $\mathbb{R}^2$ has Jordan content zero.

If $x_1 = x_2$, the line segment L is vertical. It is a subset of the boundary of the closed rectangle $[x_1, x_1 + 1] \times [y_1, y_2]$. Hence, L has Jordan content zero.

If $x_1 \neq x_2$, L is the graph of the continuous function $f : [x_1, x_2] \to \mathbb{R}$,

$$f(x) = y_1 + \frac{y_2 - y_1}{x_2 - x_1}(x - x_1).$$

Therefore L also has Jordan content zero.

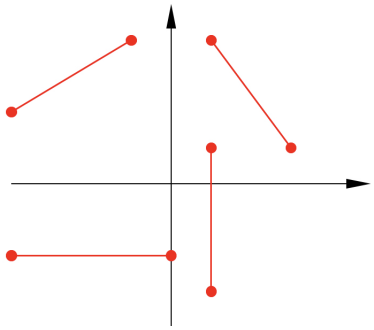


Figure 6.19: Line segments in the plane have Jordan content zero.

Example 6.30

Since the boundary of a polygon in $\mathbb{R}^2$ is a finite union of line segments, a polygon is Jordan measurable. The interior of the polygon is also Jordan measurable as it has the same boundary.

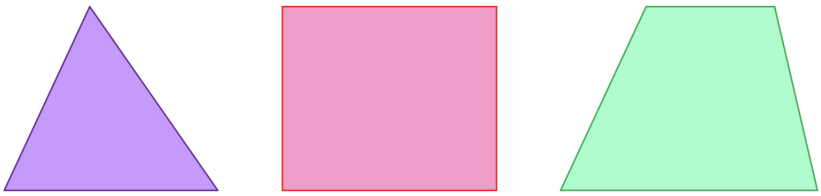


Figure 6.20: Polygons in the plane are Jordan measurable.

Example 6.31

We can argue that a line segment or the part of a plane in $\mathbb{R}^3$ contained inside a bounded set has Jordan content zero. A plane in $\mathbb{R}^3$ satisfies an equation of the form

$$ax + by + cz = d,$$

where $(a, b, c) \neq \mathbf{0}$. Therefore, one can always solve one of the variables as a function of the other two. For example, if $c \neq 0$, then

$$z = f(x, y) = \frac{d - ax - by}{c}.$$

Hence, a plane is the graph of a continuous function. If we consider the part of the plane contained within a bounded set, then it must have Jordan content zero. For example,

$$S = \{(x, y, z) \,|\, x + y + z = 3, x \geq 0, y \geq 0, z \geq 0\}$$

is the part of the plane $x + y + z = 3$ bounded inside the rectangle $[0, 3] \times [0, 3] \times [0, 3]$. Hence, S has Jordan content zero.
A line segment in $\mathbb{R}^3$ can always be regarded as a subset of a part of a plane that is contained in a bounded set. Hence, it also has Jordan content zero.

Example 6.32

The boundary of the open rectangle $U = \prod_{i=1}^{n}(a_i, b_i)$ is the same as the boundary of its closure $R = \prod_{i=1}^{n}[a_i, b_i]$. Hence, U is also a Jordan measurable set. Since U and ∂U are disjoint, and their union is R,

$$\operatorname{vol} U + \operatorname{vol} \partial U = \operatorname{vol} R.$$

Since $\operatorname{vol}(\partial U) = 0$, we have

$$\operatorname{vol} U = \operatorname{vol} R.$$

Motivated by Example 6.32, an interesting question to ask is if the subset $\mathfrak{D}$

of $\mathbb{R}^n$ is Jordan measurable, is its closure $\overline{\mathfrak{D}}$ Jordan measurable? This is answered in the following theorem.

Theorem 6.49

If $\mathfrak{D}$ is a subset of $\mathbb{R}^n$ that is Jordan measurable, so is $\overline{\mathfrak{D}}$. Moreover,

$$\operatorname{vol}\overline{\mathfrak{D}} = \operatorname{vol}\mathfrak{D}.$$

Proof

First we claim that $\partial\overline{\mathfrak{D}} \subset \partial\mathfrak{D}$. As the closure of $\mathfrak{D}$, $\overline{\mathfrak{D}}$ is a disjoint union of $\operatorname{int}\mathfrak{D}$ and $\partial\mathfrak{D}$. As the closure of $\overline{\mathfrak{D}}$, $\overline{\mathfrak{D}}$ is a disjoint union of $\operatorname{int}\overline{\mathfrak{D}}$ and $\partial\overline{\mathfrak{D}}$. Since $\mathfrak{D} \subset \overline{\mathfrak{D}}$, we have $\operatorname{int}\mathfrak{D} \subset \operatorname{int}\overline{\mathfrak{D}}$. Hence, we must have $\partial\overline{\mathfrak{D}} \subset \partial\mathfrak{D}$.
If $\mathfrak{D}$ is Jordan measurable, $\partial\mathfrak{D}$ has Jordan content zero. Since $\partial\overline{\mathfrak{D}} \subset \partial\mathfrak{D}$, $\partial\overline{\mathfrak{D}}$ also has Jordan content zero. Hence, $\overline{\mathfrak{D}}$ is Jordan measurable.
For the last statement, we use the fact that $\overline{\mathfrak{D}} = \mathfrak{D} \cup \partial\mathfrak{D}$. Notice that $\mathfrak{D} \cap \partial\mathfrak{D} \subset \partial\mathfrak{D}$. Hence, $\mathfrak{D} \cap \partial\mathfrak{D}$ has Jordan content zero. By the additivity theorem,

$$\operatorname{vol}\mathfrak{D} + \operatorname{vol}\partial\mathfrak{D} = \operatorname{vol}\overline{\mathfrak{D}}.$$

Since $\operatorname{vol}\partial\mathfrak{D} = 0$, we conclude that $\operatorname{vol}\overline{\mathfrak{D}} = \operatorname{vol}\mathfrak{D}$.

Example 6.33

Consider the set $A = (-1, 0) \cup (0, 1)$. Its closure is $\overline{A} = [-1, 1]$. Hence, $\partial\overline{A} = \{-1, 1\}$ is not equal to $\partial A = \{-1, 0, 1\}$. This also shows that even for an open set A, we does not have $\partial A = \partial\overline{A}$.

Remark 6.5

If $\mathfrak{D}$ is a bounded subset of $\mathbb{R}^n$ such that $\overline{\mathfrak{D}}$ is Jordan measurable, one cannot deduce that $\mathfrak{D}$ is Jordan measurable. An example is given by $\mathfrak{D} = [0, 1]^n \cap \mathbb{Q}^n$, which is not Jordan measurable, but $\overline{\mathfrak{D}} = [0, 1]^n$ is Jordan measurable.

Example 6.34

Let r be a positive number. We claim that the disc

$$\mathfrak{D} = \left\{(x, y) \,|\, x^2 + y^2 < r^2\right\}$$

and its closure are Jordan measurable sets. By Theorem 6.49, it is sufficient to show that $\mathfrak{D}$ is Jordan measurable. Notice that

$$\partial\mathfrak{D} = \left\{(x, y) \,|\, x^2 + y^2 = r^2\right\} = S_+ \cup S_-,$$

where

$$S_\pm = \left\{(x, y) \;\middle|\; -1 \le x \le 1, y = \pm\sqrt{r^2 - x^2}\right\}.$$

$S_\pm$ are the graphs of the functions $f_\pm : [-1, 1] \to \mathbb{R}$,

$$f_\pm(x) = \pm\sqrt{r^2 - x^2}.$$

Since $[-1, 1]$ is a Jordan measurable set in $\mathbb{R}$, and $f_\pm : [-1, 1] \to \mathbb{R}$ are continuous functions, Corollary 6.48 implies that $S_\pm = G_{f_\pm}$ have Jordan content zero. Therefore, $\mathfrak{D}$ is Jordan measurable.

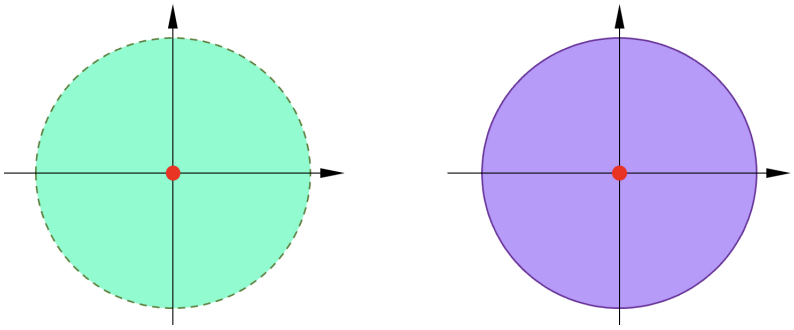


Figure 6.21: An open ball and its closure are Jordan measurable.

More generally than the open balls, we have the following.

Example 6.35

Let $[a, b]$ be a closed interval in $\mathbb{R}$, and let $f : [a, b] \to \mathbb{R}$ and $g : [a, b] \to \mathbb{R}$ be continuous functions satisfying $f(x) \leq g(x)$ for all $x \in [a, b]$. Define $\mathfrak{D}_1$ and $\mathfrak{D}_2$ to be the sets

$$\mathfrak{D}_1 = \{(x, y) \,|\, a < x < b, f(x) < y < g(x)\},$$
$$\mathfrak{D}_2 = \{(x, y) \,|\, a \leq x \leq b, f(x) \leq y \leq g(x)\}$$

Show that $\mathfrak{D}_1$ and $\mathfrak{D}_2$ are Jordan measurable sets.

Solution

Since $f : [a, b] \to \mathbb{R}$ and $g : [a, b] \to \mathbb{R}$ are continuous, $\mathfrak{D}_1$ is open and $\mathfrak{D}_2$ is closed. Since $[a, b]$ is compact, they are bounded. There is a positive number M such that

$$|f(x)| \leq M \qquad |g(x)| \leq M \qquad \text{for all } x \in [a, b].$$

Let

$$S_1 = \{(a, y) \,|\, -M \leq y \leq M\},$$
$$S_2 = \{(b, y) \,|\, -M \leq y \leq M\},$$
$$S_3 = \{(x, y) \,|\, a \leq x \leq b, y = f(x)\},$$
$$S_4 = \{(x, y) \,|\, a \leq x \leq b, y = g(x)\},$$

and let $S = S_1 \cup S_2 \cup S_3 \cup S_4$. Then

$$\mathfrak{D}_2 \setminus \mathfrak{D}_1 \subset S.$$

Since $\mathfrak{D}_1$ is open and $\mathfrak{D}_1 \subset \mathfrak{D}_2$,

$$\mathfrak{D}_1 = \operatorname{int} \mathfrak{D}_1 \subset \operatorname{int} \mathfrak{D}_2.$$

Since $\mathfrak{D}_2$ is closed and $\mathfrak{D}_1 \subset \mathfrak{D}_2$,

$$\overline{\mathfrak{D}_1} \subset \overline{\mathfrak{D}_2} = \mathfrak{D}_2.$$

Therefore,

$$\partial\mathfrak{D}_1 = \overline{\mathfrak{D}_1} \setminus \operatorname{int}\mathfrak{D}_1 \subset \mathfrak{D}_2 \setminus \mathfrak{D}_1 \subset S,$$
$$\partial\mathfrak{D}_2 = \overline{\mathfrak{D}_2} \setminus \operatorname{int}\mathfrak{D}_2 \subset \mathfrak{D}_2 \setminus \mathfrak{D}_1 \subset S.$$

Since S_1 and S_2 are line segments, they have Jordan content zero. Since S_3 and S_4 are graphs of continuous functions defined on the Jordan measurable set $[a, b]$, S_3 and S_4 also have Jordan content zero. These imply that S has Jordan content zero. Thus, $\partial\mathfrak{D}_1$ and $\partial\mathfrak{D}_2$ also have Jordan content zero, which imply that $\mathfrak{D}_1$ and $\mathfrak{D}_2$ are Jordan measurable sets.

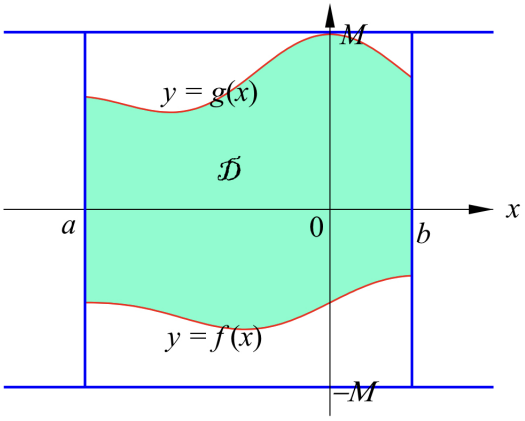


Figure 6.22: The set $\mathfrak{D} = \{(x, y) \,|\, a \le x \le b, f(x) \le y \le g(x)\}$ is Jordan measurable.

More generally than Example 6.35, we can prove the following.

Theorem 6.50

Let $\mathcal{U}$ be a Jordan measurable set in $\mathbb{R}^n$, and let $f : \mathcal{U} \to \mathbb{R}$ and $g : \mathcal{U} \to \mathbb{R}$ be bounded continuous functions on $\mathcal{U}$ satisfying $f(\mathbf{x}) \le g(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{U}$. Then the subsets

$$\mathfrak{D}_1 = \{(\mathbf{x}, y) \,|\, \mathbf{x} \in \mathcal{U}, f(\mathbf{x}) < y < g(\mathbf{x})\},$$
$$\mathfrak{D}_2 = \{(\mathbf{x}, y) \,|\, \mathbf{x} \in \mathcal{U}, f(\mathbf{x}) \le y \le g(\mathbf{x})\}$$

of $\mathbb{R}^{n+1}$ are Jordan measurable.

Sketch of Proof

Let M be a positive number such that

$$|f(\mathbf{x})| \leq M \quad \text{and} \quad |g(\mathbf{x})| \leq M \qquad \text{for all } \mathbf{x} \in \mathcal{U}.$$

The sets $\partial\mathfrak{D}_1$ and $\partial\mathfrak{D}_2$ are contained in the set $S = S_1 \cup S_2 \cup S_3$, where

$$\begin{aligned} S_1 &= \{(\mathbf{x}, y) \,|\, \mathbf{x} \in \partial\mathcal{U}, -M \leq y \leq M\}, \\ S_2 &= \{(\mathbf{x}, y) \,|\, \mathbf{x} \in \mathcal{U}, y = f(\mathbf{x})\}, \\ S_3 &= \{(\mathbf{x}, y) \,|\, \mathbf{x} \in \mathcal{U}, y = g(\mathbf{x})\}. \end{aligned}$$

The sets S_1, S_2 and S_3 have Jordan content zero.

Example 6.36

We claim that an open ball $B(\mathbf{x}_0, r)$ in $\mathbb{R}^n$ and its closure are Jordan measurable sets. It is sufficient to consider the case where $\mathbf{x}_0 = \mathbf{0}$ and $r = 1$. Let $B^n = B(\mathbf{0}, 1)$. We will show that B^n is Jordan measurable by induction on n. Then $\overline{B^n}$ is also Jordan measurable.
When $n = 1$, $B^1 = (-1, 1)$ is an interval whose boundary is the two point set $\{-1, 1\}$ which has Jordan content zero. For $n \geq 1$, assume that B^n is a Jordan measurable subset of $\mathbb{R}^n$. Notice that

$$B^{n+1} = \{(\mathbf{x}, y) \,|\, \mathbf{x} \in B^n, f_-(\mathbf{x}) < y < f_+(\mathbf{x})\},$$

where $f_\pm : B^n \to \mathbb{R}$ are bounded continuous functions defined by

$$f_\pm(x_1, \ldots, x_n) = \pm\sqrt{1 - x_1^2 - \cdots - x_n^2}.$$

By inductive hypothesis, B^n is Jordan measurable. By Theorem 6.50, B^{n+1} is also Jordan measurable.

Now we give some examples of Riemann integrable functions.

Example 6.37

Let $\mathfrak{D} = \{(x, y, z) \,|\, x^2 + y^2 < 4, -3 \leq z \leq 3\}$, and let $f : \mathfrak{D} \to \mathbb{R}$ be the function defined as

$$f(x, y, z) = x^2 + 4y^2 + 9z^2.$$

Explain why $f : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable.

Solution

The set $\mathcal{U} = \{(x, y, z) \,|\, x^2 + y^2 < 4\}$ is an open ball. Hence, it is Jordan measurable. The functions $g_\pm : \mathcal{U} \to \mathbb{R}$, $g_\pm(x, y) = \pm 3$ are continuous functions. Theorem 6.50 implies that $\mathfrak{D}$ is Jordan measurable. The function $f(x, y, z) = x^2 + 4y^2 + 9z^2$ is a polynomial. Hence, it is continuous on $\overline{\mathfrak{D}}$, and hence, it is bounded and continuous on $\mathfrak{D}$. Therefore, $f : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable.

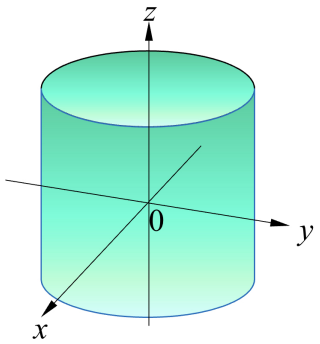


Figure 6.23: The Jordan measurable set in Example 6.37.

Example 6.38

Let $\mathfrak{D} = \{(x, y) \,|\, x^2 + y^2 < 1\}$, and let $f : \mathfrak{D} \to \mathbb{R}$ be the function

$$f(x, y) = \begin{cases} x^2, & \text{if } x < y, \\ y^2 + 1, & \text{if } x \geq y. \end{cases}$$

Explain why $f : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable.

Solution

The set $\mathfrak{D}$ is an open ball. Hence, it is Jordan measurable. The set of discontinuities of the function $f : \mathfrak{D} \to \mathbb{R}$ is contained in the line segment L from the point $(-1, -1)$ to the point $(1, 1)$. Since L has Jordan content zero, $f : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable.

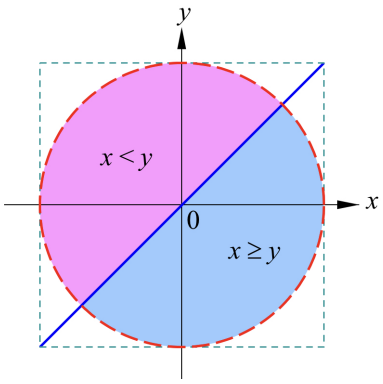


Figure 6.24: The Jordan measurable set in Example 6.38.

At the end of this section, let us prove a few interesting theorems. The next theorem shows that a set with Jordan content zero cannot have nonempty interior.

Theorem 6.51

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$ that has Jordan content zero. Then $\operatorname{int} \mathfrak{D} = \emptyset$.

Proof

We use proof by contradiction. If $\operatorname{int} \mathfrak{D} \neq \emptyset$, it is an open set that contains at least one point $\mathbf{u} = (u_1, \ldots, u_n)$. By definition of interior points, there exists $r > 0$ such that $B(\mathbf{u}, r) \subset \mathfrak{D}$. There exists $\delta > 0$ such that the rectangle $\mathbf{I}_\delta = \prod_{i=1}^{n} [u_i - \delta, u_i + \delta]$ is contained in $B(\mathbf{u}, r)$, and hence in $\mathfrak{D}$. Since $\mathfrak{D}$ has Jordan content zero, we find that $\mathbf{I}_\delta$ also has Jordan content zero. But

$$\operatorname{vol}(\mathbf{I}_\delta) = (2\delta)^n > 0.$$

This gives a contradiction.

The next one is the mean value theorem for integrals.

Theorem 6.52 Mean Value Theorem for Integrals

Let $\mathfrak{D}$ be a closed and bounded Jordan measurable set in $\mathbb{R}^n$, and let $f : \mathfrak{D} \to \mathbb{R}$ be a continuous function. If $\mathfrak{D}$ is connected or path-connected, then there is a point $\mathbf{x}_0$ in $\mathfrak{D}$ such that

$$\int_{\mathfrak{D}} f(\mathbf{x}) d\mathbf{x} = f(\mathbf{x}_0) \operatorname{vol}(\mathfrak{D}).$$

Proof

Since $\mathfrak{D}$ is compact and $f : \mathfrak{D} \to \mathbb{R}$ is continuous, extreme value theorem asserts that there exist points $\mathbf{u}$ and $\mathbf{v}$ in $\mathfrak{D}$ such that

$$f(\mathbf{u}) \leq f(\mathbf{x}) \leq f(\mathbf{v}) \qquad \text{for all } \mathbf{x} \in \mathfrak{D}.$$

Since $\mathfrak{D}$ is a Jordan measurable set and $f : \mathfrak{D} \to \mathbb{R}$ is continuous, $f : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable. The monotonicity theorem implies that

$$f(\mathbf{u}) \operatorname{vol}(\mathfrak{D}) \leq \int_{\mathfrak{D}} f(\mathbf{x}) d\mathbf{x} \leq f(\mathbf{v}) \operatorname{vol}(\mathfrak{D}).$$

If $\operatorname{vol}(\mathfrak{D}) = 0$, we can take $\mathbf{x}_0$ to be any point in $\mathfrak{D}$. If $\operatorname{vol}(\mathfrak{D}) \neq 0$, notice that

$$c = \frac{1}{\operatorname{vol}(\mathfrak{D})} \int_{\mathfrak{D}} f(\mathbf{x}) d\mathbf{x}$$

satisfies

$$f(\mathbf{u}) \leq c \leq f(\mathbf{v}).$$

Since $\mathfrak{D}$ is connected or path-connected, and $f : \mathfrak{D} \to \mathbb{R}$ is continuous, intermediate value theorem asserts that $f(\mathfrak{D})$ must be an interval. Since $f(\mathbf{u})$ and $f(\mathbf{v})$ are in $f(\mathfrak{D})$ and c is in between them, c must also be in $f(\mathfrak{D})$. This means that there is an $\mathbf{x}_0$ in $\mathfrak{D}$ such that

$$\frac{1}{\operatorname{vol}(\mathfrak{D})} \int_{\mathfrak{D}} f(\mathbf{x}) d\mathbf{x} = c = f(\mathbf{x}_0).$$

Exercises 6.3

Question 1

Explain why the set

$$\mathfrak{D} = \left\{(x, y, z) \,|\, 4x^2 + y^2 + 9z^2 < 36\right\}$$

is Jordan measurable.

Question 2

Explain why the set

$$\mathfrak{D} = \{(x, y, z) \,|\, x \geq 0, y \geq 0, z \geq 0, 3x + 4y + 6z \leq 12\}$$

is Jordan measurable.

Question 3

Let $\mathfrak{D} = \left\{(x, y, z) \,|\, x^2 + y^2 + z^2 = 25\right\}$, and let $f : \mathfrak{D} \to \mathbb{R}$ be the function defined as

$$f(x, y, z) = \begin{cases} 1, & \text{if } x, y \text{ and } z \text{ are rational,} \\ 0, & \text{otherwise.} \end{cases}$$

Explain why $f : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable and find $\displaystyle\int_{\mathfrak{D}} f$.

Question 4

Let $\mathfrak{D} = \left\{(x, y, z) \,|\, x^2 + y^2 + z^2 \leq 25\right\}$, and let $f : \mathfrak{D} \to \mathbb{R}$ be the function defined as

$$f(x, y, z) = ze^{|xy|}.$$

Explain why $f : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable.

Question 5

Let $\mathfrak{D} = [0, 2] \times (-2, 5) \times (1, 7]$, and let $f : \mathfrak{D} \to \mathbb{R}$ be the function defined as

$$f(x, y, z) = \begin{cases} x + y, & \text{if } x < y + z, \\ 2x - y, & \text{if } x \geq y + z. \end{cases}$$

Explain why $f : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable.

Question 6

Let $\mathfrak{D} = \left\{(x, y, z) \,|\, 4x^2 + 9y^2 \leq 36, 0 \leq z \leq x^2 + y^2\right\}$, and let $f : \mathfrak{D} \to \mathbb{R}$ be the function defined as

$$f(x, y, z) = \begin{cases} x, & \text{if } x < y + z, \\ y + z, & \text{if } x \geq y + z. \end{cases}$$

Explain why $f : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable.

Question 7

If $\mathfrak{D}$ is a Jordan measurable set that is contained in the closed rectangle $\mathbf{I}$, show that $\mathbf{I} \setminus \mathfrak{D}$ is also Jordan measurable. Moreover,

$$\operatorname{vol}(\mathbf{I} \setminus \mathfrak{D}) = \operatorname{vol}(\mathbf{I}) - \operatorname{vol}(\mathfrak{D}).$$

Question 8

If $\mathfrak{D}_1$ and $\mathfrak{D}_2$ are Jordan measurable sets and $\mathfrak{D}_2$ is contained in $\mathfrak{D}_1$, show that $\mathfrak{D}_1 \setminus \mathfrak{D}_2$ is also Jordan measurable. Moreover,

$$\operatorname{vol}(\mathfrak{D}_1 \setminus \mathfrak{D}_2) = \operatorname{vol}(\mathfrak{D}_1) - \operatorname{vol}(\mathfrak{D}_2).$$

Question 9

If $\mathfrak{D}$ is a Jordan measurable set, show that $\operatorname{int} \mathfrak{D}$ is also Jordan measurable. Moreover,

$$\operatorname{vol}(\operatorname{int} \mathfrak{D}) = \operatorname{vol}(\mathfrak{D}).$$

Question 10

Let $\mathfrak{D}_1$ and $\mathfrak{D}_2$ be Jordan measurable sets in $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively. Assume that $\mathfrak{D}_1$ has Jordan content zero. Show that the set $\mathfrak{D} = \mathfrak{D}_1 \times \mathfrak{D}_2$ in $\mathbb{R}^{m+n}$ also has Jordan content zero.

Question 11

Let $\mathfrak{D} = \{(x, y) \,|\, x^2 + y^2 \leq 9\}$. Show that the integral $\int_{\mathfrak{D}} x dx dy$ exist and is equal to 0.

Question 12

let $\mathfrak{D}$ be a Jordan measurable set, and let $f : \mathfrak{D} \to \mathbb{R}$ be a Riemann integrable function. If $g : \overline{\mathfrak{D}} \to \mathbb{R}$ is a bounded function such that $g(\mathbf{x}) = f(\mathbf{x})$ for all $\mathbf{x}$ in $\mathfrak{D}$, show that $g : \overline{\mathfrak{D}} \to \mathbb{R}$ is Riemann integrable and

$$\int_{\overline{\mathfrak{D}}} g = \int_{\mathfrak{D}} f.$$

6.4 Iterated Integrals and Fubini's Theorem

In Section 6.3, we have given a sufficient condition for a function $f : \mathfrak{D} \to \mathbb{R}$ to be Riemann integrable.

Riemann Integrable Functions

If $\mathfrak{D}$ is a subset of $\mathbb{R}^n$ such that a constant function on $\mathfrak{D}$ is Riemann integrable, then any bounded function on $\mathfrak{D}$ whose set of discontinuities is a set that has Jordan content zero is Riemann integrable.

However, we have not discussed any strategy to compute a Riemann integral, except by using a sequence of partitions $\{\mathbf{P}_k\}$ with

$$\lim_{k\to\infty} |\mathbf{P}_k| = 0.$$

This is a practical approach if one has a computer, but it is not feasible for hand calculations. Besides, it might also be difficult for us to understand the dependence of the integral on the parameters in the integrand. When $n = 1$, we have seen that the fundamental theorem of calculus gives us a powerful tool to calculate a Riemann integral when the integrand is a continuous function that has explicit antiderivatives. To be able to apply this powerful tool in the multivariable context, we need to relate multiple integrals with iterated integrals. This is the topic that is studied in this section.

As a motivation, consider a continuous function $f : [a, b] \times [c, d] \to \mathbb{R}$ defined on the closed rectangle $\mathbf{I} = [a, b] \times [c, d]$ in $\mathbb{R}^2$. If $\mathbf{P} = (P_1, P_2)$ is a partition of $\mathbf{I}$ with

$$P_1 = \{x_0, x_1, \ldots, x_k\} \quad \text{and} \quad P_2 = \{y_0, y_1, \ldots, y_l\},$$

there are kl rectangles in the partition $\mathbf{P}$. Denote the rectangles by $\mathbf{J}_{i,j}$ with $1 \leq i \leq k$ and $1 \leq j \leq l$, where

$$\mathbf{J}_{i,j} = [x_{i-1}, x_i] \times [y_{j-1}, y_j].$$

Choose a set of intermediate points $A = \{\alpha_i\}$ for the partition P_1, and a set of intermediate points $B = \{\beta_j\}$ for the partition P_2. Let

$$\boldsymbol{\xi}_{i,j} = (\alpha_i, \beta_j) \qquad \text{for } 1 \leq i \leq k, 1 \leq j \leq l.$$

Then $C = \left\{ \boldsymbol{\xi}_{i,j} \,|\, 1 \le i \le k, 1 \le j \le l \right\}$ is a choice of intermediate points for the partition $\mathbf{P}$. The Riemann sum $R(f, \mathbf{P}, C)$ is given by

$$R(f, \mathbf{P}, C) = \sum_{i=1}^{k} \sum_{j=1}^{l} f(\alpha_i, \beta_j)(x_i - x_{i-1})(y_j - y_{j-1}). \tag{6.4}$$

Since it is a finite sum, it does not matter which order we perform the summation. For fixed $x \in [a, b]$, let $g_x : [c, d] \to \mathbb{R}$ be the function

$$g_x(y) = f(x, y), \qquad y \in [c, d].$$

If we perform the sum over j in (6.4) first, we find that

$$\begin{aligned} R(f, \mathbf{P}, C) &= \sum_{i=1}^{k} \left(\sum_{j=1}^{l} g_{\alpha_i}(\beta_j)(y_j - y_{j-1}) \right) (x_i - x_{i-1}) \\ &= \sum_{i=1}^{k} R(g_{\alpha_i}, P_2, B)(x_i - x_{i-1}). \end{aligned}$$

Since $g_{\alpha_i} : [c, d] \to \mathbb{R}$ is continuous, it is Riemann integrable. Therefore,

$$\lim_{|P_2| \to 0} R(g_{\alpha_i}, P_2, B) = \int_c^d g_{\alpha_i}(y) dy.$$

This prompts us to define the function $F : [a, b] \to \mathbb{R}$ by

$$F(x) = \int_c^d g_x(y) dy = \int_c^d f(x, y) dy.$$

Then

$$\begin{aligned} \lim_{|P_2| \to 0} R(f, \mathbf{P}, C) &= \sum_{i=1}^{k} \lim_{|P_2| \to 0} R(g_{\alpha_i}, P_2, B)(x_i - x_{i-1}) \\ &= \sum_{i=1}^{k} F(\alpha_i)(x_i - x_{i-1}) = R(F, P_1, A). \end{aligned}$$

If $F : [a, b] \to \mathbb{R}$ is also Riemann integrable, we would have

$$\begin{aligned} \lim_{|P_1| \to 0} \lim_{|P_2| \to 0} R(f, \mathbf{P}, C) &= \lim_{|P_1| \to 0} R(F, P_1, A) \\ &= \int_a^b F(x) dx = \int_a^b \left(\int_c^d f(x, y) dy \right) dx. \end{aligned}$$

Interchanging the roles of of x and y, or equivalently, summing over i first in (6.4), we find that

$$\lim_{|P_2|\to 0}\lim_{|P_1|\to 0} R(f,\mathbf{P},C) = \int_c^d \left(\int_a^b f(x,y)dx\right) dy.$$

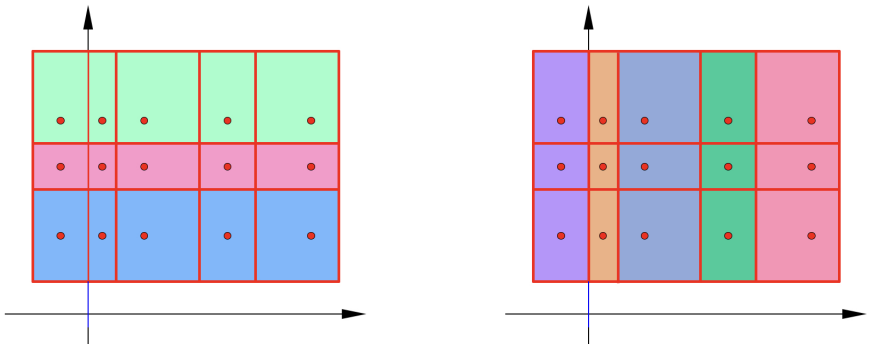


Figure 6.25: Given a partition $\mathbf{P}$ of a rectangle, one can sum over the rectangles row by row, or column by column.

Since

$$|\mathbf{P}| = \sqrt{|P_1|^2 + |P_2|^2},$$

$|\mathbf{P}| \to 0$ if and only if $(|P_1|, |P_2|) \to (0,0)$. The question becomes whether the two limits

$$\lim_{|P_1|\to 0}\lim_{|P_2|\to 0} R(f,\mathbf{P},C) \quad \text{and} \quad \lim_{|P_2|\to 0}\lim_{|P_1|\to 0} R(f,\mathbf{P},C)$$

are equal; and whether they are equal to the limit

$$\lim_{(|P_1|,|P_2|)\to(0,0)} R(f,\mathbf{P},C).$$

Remark 6.6

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function defined as

$$f(x,y) = \frac{x^2}{x^2+y^2}, \qquad (x,y) \in \mathbb{R}^2 \setminus \{(0,0)\}.$$

We find that

$$\lim_{y\to 0}\lim_{x\to 0} f(x,y) = \lim_{y\to 0} 0 = 0,$$
$$\lim_{x\to 0}\lim_{y\to 0} f(x,y) = \lim_{x\to 0} 1 = 1.$$

Hence,

$$\lim_{y\to 0}\lim_{x\to 0} f(x,y) \neq \lim_{x\to 0}\lim_{y\to 0} f(x,y).$$

This example shows that we cannot simply interchange the order of limits. In fact, the limit

$$\lim_{(x,y)\to(0,0)} f(x,y)$$

does not exist.

The integrals

$$\int_c^d \left(\int_a^b f(x,y)dx\right) dy \quad \text{and} \quad \int_a^b \left(\int_c^d f(x,y)dy\right) dx$$

are called *iterated integrals*.

Definition 6.17 Iterated Integrals

Let n be a positive integer larger than 1, and let k be a positive integer less than n. Denote a point in $\mathbb{R}^n$ by $(\mathbf{x},\mathbf{y})$, where $\mathbf{x}\in\mathbb{R}^k$ and $\mathbf{y}\in\mathbb{R}^{n-k}$. Given that $\mathbf{I}=\prod_{i=1}^{n}[a_i,b_i]$ is a closed rectangle in $\mathbb{R}^n$, let

$$\mathbf{I_x} = \prod_{i=1}^{k}[a_i,b_i] \quad \text{and} \quad \mathbf{I_y} = \prod_{i=k+1}^{n}[a_i,b_i].$$

If $f:\mathbf{I}\to\mathbb{R}$ is a bounded function defined on $\mathbf{I}$, an interated integral is an integral of the form

$$\int_{\mathbf{I_y}}\int_{\mathbf{I_x}} f(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y} \quad \text{or} \quad \int_{\mathbf{I_x}}\int_{\mathbf{I_y}} f(\mathbf{x},\mathbf{y})d\mathbf{y}d\mathbf{x},$$

whenever they exist.

Let us consider the following example.

Example 6.39

Let $g : [a, b] \to \mathbb{R}$ and $h : [a, b] \to \mathbb{R}$ be continuous functions defined on $[a, b]$ such that $g(x) \leq h(x)$ for all $x \in [a, b]$. Consider the set $\mathfrak{D}$ defined as

$$\mathfrak{D} = \{(x, y) \,|\, a \leq x \leq b, g(x) \leq y \leq h(x)\}.$$

Let $\chi_{\mathfrak{D}} : \mathbb{R}^2 \to \mathbb{R}$ be the corresponding characteristic function. If

$$c \leq g(x) \leq h(x) \leq d \qquad \text{for all } \; x \in [a, b],$$

then $\mathbf{I} = [a, b] \times [c, d]$ is a closed rectangle that contains $\mathfrak{D}$. We have seen that $\mathfrak{D}$ is a Jordan measurable set. Hence, $\chi_{\mathfrak{D}} : \mathbf{I} \to \mathbb{R}$ is a Riemann integrable function. For any $x \in [a, b]$,

$$\int_c^d \chi_{\mathfrak{D}}(x, y)dy = \int_{g(x)}^{h(x)} dy = h(x) - g(x).$$

Therefore, the iterated integral $\displaystyle\int_a^b \left(\int_c^d \chi_{\mathfrak{D}}(x, y)dy \right) dx$ is equal to

$$\int_a^b \left(\int_c^d \chi_{\mathfrak{D}}(x, y)dy \right) dx = \int_a^b (h(x) - g(x))\, dx.$$

In single variable calculus, we have learned that the integral $\displaystyle\int_a^b (h(x) - g(x))\, dx$ gives the area of $\mathfrak{D}$. Thus, in this case, we have

$$\int_a^b \left(\int_c^d \chi_{\mathfrak{D}}(x, y)dy \right) dx = \text{vol}\,(\mathfrak{D}) = \int_{[a,b]\times[c,d]} \chi_{\mathfrak{D}}(x, y)dxdy.$$

Namely, the iterated integral is equal to the double integral.

The following theorem is the general case when $n = 2$.

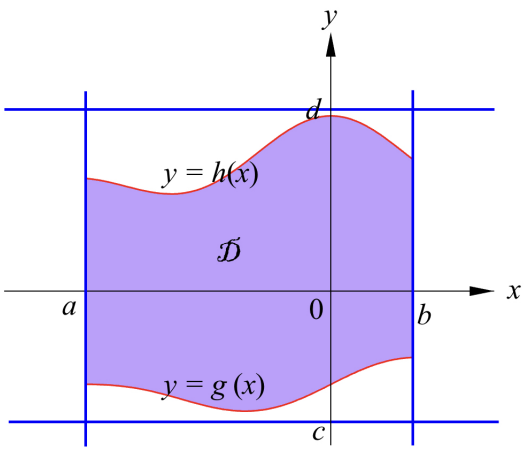


Figure 6.26: The domain $\mathfrak{D} = \{(x, y) \mid a \leq x \leq b, g(x) \leq y \leq h(x)\}$.

Theorem 6.53 Fubini's Theorem in the Plane

Let $\mathbf{I} = [a, b] \times [c, d]$, and let $f : \mathbf{I} \to \mathbb{R}$ be a Riemann integrable function. For each $x \in [a, b]$, define the function $g_x : [c, d] \to \mathbb{R}$ by

$$g_x(y) = f(x, y), \quad y \in [c, d].$$

If $g_x : [c, d] \to \mathbb{R}$ is Riemann integrable for each $x \in [a, b]$, let $F : [a, b] \to \mathbb{R}$ be the function defined as

$$F(x) = \int_c^d g_x(y)dy = \int_c^d f(x, y)dy.$$

Then we have the following.

(a) The function $F : [a, b] \to \mathbb{R}$ is Riemann integrable.

(b) The integral of $f : [a, b] \times [c, d] \to \mathbb{R}$ is equal to the integral of $F : [a, b] \to \mathbb{R}$. Namely,

$$\int_{[a,b]\times[c,d]} f = \int_a^b F.$$

Equivalently,

$$\int_{[a,b]\times[c,d]} f(x, y)dxdy = \int_a^b \int_c^d f(x, y)dydx.$$

Proof

Let

$$I = \int_{[a,b]\times[c,d]} f(x,y)dxdy.$$

We will show that for any $\varepsilon > 0$, there exists $\delta > 0$ such that if P is a partition of $[a,b]$ with $|P| < \delta$, and $A = \{\alpha_i\}$ is any set of intermediate points for P, then

$$|R(F,P,A) - I| < \varepsilon.$$

This will prove both (a) and (b).

Fixed $\varepsilon > 0$. Since $f : [a,b] \times [c,d] \to \mathbb{R}$ is Riemann integrable with integral I, there exists $\delta_0 > 0$ such that if $\mathbf{P} = (P_1, P_2)$ is a partition of $[a,b] \times [c,d]$ with $|\mathbf{P}| < \delta_0$, then

$$U(f,\mathbf{P}) - L(f,\mathbf{P}) < \varepsilon.$$

Take $\delta = \delta_0/2$. Let $P = \{x_0, x_1, \ldots, x_k\}$ be a partition of $[a,b]$ with $|P| < \delta$. Take any partition $P_2 = \{y_0, y_1, \ldots, y_l\}$ of $[c,d]$ such that $|P_2| < \delta$. Let $\mathbf{P} = (P_1, P_2)$, where $P_1 = P$. Then $|\mathbf{P}| < \sqrt{2}\delta < \delta_0$. For $1 \le i \le k, 1 \le j \le l$, let

$$m_{i,j} = \inf_{(x,y)\in[x_{i-1},x_i]\times[y_{j-1},y_j]} f(x,y),$$
$$M_{i,j} = \sup_{(x,y)\in[x_{i-1},x_i]\times[y_{j-1},y_j]} f(x,y).$$

Then

$$L(f,\mathbf{P}) = \sum_{i=1}^{k}\sum_{j=1}^{l} m_{i,j}(x_i - x_{i-1})(y_j - y_{j-1}),$$
$$U(f,\mathbf{P}) = \sum_{i=1}^{k}\sum_{j=1}^{l} M_{i,j}(x_i - x_{i-1})(y_j - y_{j-1}).$$

Now let $A = \{\alpha_i\}$ be any choice of intermediate points for the partition $P = P_1$. Notice that for any $1 \le i \le k$, additivity theorem says that

$$F(\alpha_i) = \sum_{j=1}^{l}\int_{y_{j-1}}^{y_j} g_{\alpha_i}(y)dy = \sum_{j=1}^{l}\int_{y_{j-1}}^{y_j} f(\alpha_i, y)dy.$$

Since

$$m_{i,j} \leq f(\alpha_i, y) \leq M_{i,j} \qquad \text{for all } y \in [y_{j-1}, y_j],$$

we find that for $1 \leq j \leq l$,

$$m_{i,j}(y_i - y_{j-1}) \leq \int_{y_{j-1}}^{y_j} f(\alpha_i, y)dy \leq M_{i,j}(y_i - y_{j-1}).$$

It follows that

$$\sum_{j=1}^{l} m_{i,j}(y_j - y_{j-1}) \leq F(\alpha_i) \leq \sum_{j=1}^{l} M_{i,j}(y_j - y_{j-1}).$$

Multiply by $(x_i - x_{i-1})$, and sum over i from 1 to k, we find that

$$L(f, \mathbf{P}) \leq \sum_{i=1}^{k} F(\alpha_i)(x_i - x_{i-1}) \leq U(f, \mathbf{P}).$$

In other words,

$$L(f, \mathbf{P}) \leq R(F, P, A) \leq U(f, \mathbf{P}).$$

Since we also have

$$L(f, \mathbf{P}) \leq I \leq U(f, \mathbf{P}),$$

we find that

$$|R(F, P, A) - I| \leq U(f, \mathbf{P}) - L(f, \mathbf{P}) < \varepsilon.$$

This completes the proof.

Example 6.40

Evaluate the integral $\int_{[0,1]\times[0,1]} x\sin(xy)dxdy$.

Solution

The function $f : [0, 1] \times [0, 1] \to \mathbb{R}$, $f(x, y) = x\sin(xy)$ is a continuous function. Hence, it is Riemann integrable.

For each $x \in [0,1]$, the function $g_x : [0,1] \to \mathbb{R}$, $g_x(y) = x\sin(xy)$ is also continuous. Hence, $g_x : [0,1] \to \mathbb{R}$ is Riemann integrable. By Fubini's theorem,

$$\begin{aligned}\int_{[0,1]\times[0,1]} x\sin(xy)dxdy &= \int_0^1 \int_0^1 x\sin(xy)dydx \\ &= \int_0^1 [-\cos(xy)]_{y=0}^{y=1}\, dx \\ &= \int_0^1 (1-\cos x)\, dx \\ &= 1 - [\sin x]_0^1 = 1 - \sin 1.\end{aligned}$$

The roles of x and y in Fubini's theorem can be interchanged, and we obain the following.

Corollary 6.54

Assume that $f : [a,b] \times [c,d] \to \mathbb{R}$ is a Riemann integrable function such that for each $x \in [a,b]$, the function $g_x : [c,d] \to \mathbb{R}$, $g_x(y) = f(x,y)$ is Riemann integrable; and for each $y \in [c,d]$, the function $h_y : [a,b] \to \mathbb{R}$, $h_y(x) = f(x,y)$ is Riemann integrable. Then we can interchange the order of integration. Namely,

$$\int_c^d \int_a^b f(x,y)dxdy = \int_a^b \int_c^d f(x,y)dydx.$$

Example 6.41

If we evaluate the iterated integral $\int_0^1 \int_0^1 x\sin(xy)dxdy$ directly, it would be quite tedious as we need to apply integration by parts to evaluate the integral $\int_0^1 x\sin(xy)dx$. Using Corollary 6.54, we can interchange the order of integration and obtain

$$\int_0^1 \int_0^1 x\sin(xy)dxdy = \int_0^1 \int_0^1 x\sin(xy)dydx = 1 - \sin 1.$$

Remark 6.7

The assumption that $f : [a, b]\times[c, d] \to \mathbb{R}$ is Riemann integrable is essential in Fubini's theorem. It does not follow from the fact that for each $x \in [a, b]$, the function $g_x : [c, d] \to \mathbb{R}$ is Riemann integrable, and the function $F : [a, b] \to \mathbb{R}$,

$$F(x) = \int_c^d g_x(y)dy$$

is Riemann integrable. For example, let $g : [-1, 1] \to \mathbb{R}$ and $h : [-1, 1] \to \mathbb{R}$ be the functions defined as

$$g(x) = \begin{cases} 1, & \text{if } x \text{ is rational,} \\ -1, & \text{if } x \text{ is irrational,} \end{cases} \qquad h(y) = \begin{cases} 1, & \text{if } y \geq 0, \\ -1, & \text{if } y < 0. \end{cases}.$$

Then define the function $f : [-1, 1] \times [-1, 1] \to \mathbb{R}$ by

$$f(x, y) = g(x)h(y).$$

Since $h : [-1, 1] \to \mathbb{R}$ is a step function, it is Riemann integrable and

$$\int_{-1}^1 h(y)dy = \int_{-1}^0 h(y)dy + \int_0^1 h(y)dy = 0.$$

Hence, for fixed $x \in [-1, 1]$,

$$\int_{-1}^1 f(x, y)dy = 0.$$

Thus, the function $F : [-1, 1] \to \mathbb{R}$,

$$F(x) = \int_{-1}^1 f(x, y)dy,$$

being a function that is always zero, is Riemann integrable with integral 0. It follows that

$$\int_{-1}^1 \int_{-1}^1 f(x, y)dydx = 0.$$

However, one can prove that the function $f : [-1, 1] \times [-1, 1] \to \mathbb{R}$ is not Riemann integrable, using the same way that we show that a Dirichlet's function is not Riemann integrable.

The fact that for each $x \in [a,b]$, the function $g_x : [c,d] \to \mathbb{R}$, $g_x(y) = f(x,y)$ is Riemann integrable also does not follow from the fact that $f : [a,b]\times[c,d] \to \mathbb{R}$ is Riemann integrable. Consider for example the function $f : [-1,1] \times [0,1] \to \mathbb{R}$,

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

Then the set of discontinuities $\mathcal{N}$ of $f : [-1,1] \times [0,1] \to \mathbb{R}$ is the line segment between the point $(0,0)$ and the point $(0,1)$. Hence, $\mathcal{N}$ has Jordan content 0. Therefore, $f : [-1,1] \times [0,1] \to \mathbb{R}$ is Riemann integrable. For $x = 0$, $g_0 : [0,1] \to \mathbb{R}$ is the Dirichlet's function. Hence, $g_0 : [0,1] \to \mathbb{R}$ is not Riemann integrable.

Now we consider the case depicted in Example 6.39 for more general functions.

Theorem 6.55

Let $g : [a,b] \to \mathbb{R}$ and $h : [a,b] \to \mathbb{R}$ be continuous functions defined on $[a,b]$ such that $g(x) \le h(x)$ for all $x \in [a,b]$, and let $\mathfrak{D}$ be the set

$$\mathfrak{D} = \{(x,y) \,|\, a \le x \le b, g(x) \le y \le h(x)\}.$$

If $f : \mathfrak{D} \to \mathbb{R}$ is a continuous function, then

$$\int_{\mathfrak{D}} f(x,y)dxdy = \int_a^b \int_{g(x)}^{h(x)} f(x,y)dydx.$$

Proof

Since g and h are continuous functions, there exist numbers c and d such that

$$c \le g(x) \le h(x) \le d \qquad \text{for all } x \in [a,b].$$

Then $\mathbf{I} = [a,b] \times [c,d]$ be a closed rectangle that contains $\mathfrak{D}$. We have shown before that $\mathfrak{D}$ is a Jordan measurable set and $f : \mathfrak{D} \to \mathbb{R}$ is Riemann integrable. Therefore, $\check{f} : \mathbf{I} \to \mathbb{R}$ is Riemann integrable.

On the other hand, for each $x \in [a,b]$, the function $g_x : [c,d] \to \mathbb{R}$ is a piecewise continuous function given by

$$g_x(y) = \begin{cases} 0, & \text{if } c \le y < g(x), \\ f(x,y), & \text{if } g(x) \le y \le h(x), \\ 0, & \text{if } h(x) < y \le d. \end{cases}$$

Hence, $g_x : [c,d] \to \mathbb{R}$ is Riemann integrable and for $x \in [a,b]$,

$$F(x) = \int_c^d g_x(y)dy = \int_{g(x)}^{h(x)} f(x,y)dy.$$

By Fubini's theorem in the plane, the function $F : [a,b] \to \mathbb{R}$ is Riemann integrable, and

$$\int_a^b \int_{g(x)}^{h(x)} f(x,y)dydx = \int_a^b F(x)dx = \int_{\mathbf{I}} \check{f}(x,y)dxdy = \int_{\mathfrak{D}} f(x,y)dxdy.$$

Again, the roles of x and y in Theorem 6.55 can be interchanged. Let us look at the following example.

Example 6.42

Let $\mathfrak{D}$ be the region in the plane bounded between the curve $y^2 = x$ and the line L between the points $(1,1)$ and $(4,-2)$. Evaluate the integral $\int_{\mathfrak{D}} ydxdy$.

Solution

The equation of the line L is $x + y = 2$. Hence,

$$\mathfrak{D} = \left\{(x,y) \mid -2 \le y \le 1, y^2 \le x \le 2 - y\right\}.$$

Using Fibini's theorem, we find that

$$\int_{\mathfrak{D}} y dx dy = \int_{-2}^{1}\int_{y^2}^{2-y} y dx dy = \int_{-2}^{1} y(2-y-y^2)dy$$
$$= \int_{-2}^{1}(2y - y^2 - y^3)dy = \left[y^2 - \frac{y^3}{3} - \frac{y^4}{4}\right]_{-2}^{1} = -\frac{9}{4}.$$

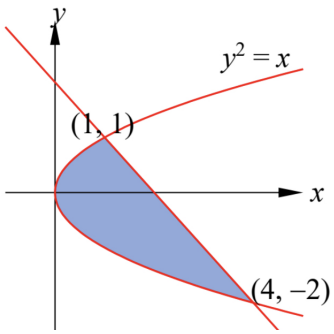


Figure 6.27: The domain $\mathfrak{D} = \{(x, y) \mid -2 \leq y \leq 1, y^2 \leq x \leq 2 - y\}$.

In Example 6.42, it will be harder if one prefers to integrate over y first.

Example 6.43

Let

$$\mathfrak{D} = \left\{(x, y) \,|\, x \geq 0, y \geq 0, 4x^2 + 9y^2 \leq 36\right\}.$$

Evaluate the integral $\displaystyle\int_{\mathfrak{D}} x dx dy$.

Solution

The set $\mathfrak{D}$ can be expressed in two different ways.

$$\mathfrak{D} = \left\{(x, y) \;\middle|\; 0 \leq x \leq 3, 0 \leq y \leq \frac{1}{3}\sqrt{36 - 4x^2}\right\},$$
$$\mathfrak{D} = \left\{(x, y) \;\middle|\; 0 \leq y \leq 2, 0 \leq x \leq \frac{1}{2}\sqrt{36 - 9y^2}\right\}.$$

The function $f : \mathfrak{D} \to \mathbb{R}$, $f(x, y) = x$ is continuous. Hence, the integral $\int_{\mathfrak{D}} x dx dy$ is equal to iterated integrals, which we can integrate with respect to x first, or with respect to y first. If we integrate with respect to y first, we find that

$$\int_{\mathfrak{D}} x dx dy = \int_0^3 \int_0^{\frac{1}{3}\sqrt{36-4x^2}} x dy dx = \frac{1}{3} \int_0^3 x\sqrt{36-4x^2} dx.$$

This integral needs to be computed using integration by substitution. If we integrate over x first, we find that

$$\begin{aligned} \int_{\mathfrak{D}} x dx dy &= \int_0^2 \int_0^{\frac{1}{2}\sqrt{36-9y^2}} x dx dy \\ &= \frac{1}{2} \left[x^2\right]_{y=0}^{y=\frac{1}{2}\sqrt{36-9y^2}} = \frac{1}{8} \int_0^2 (36 - 9y^2) dy. \end{aligned}$$

This integral can be easily evaluated to give

$$\int_{\mathfrak{D}} x dx dy = \frac{1}{8} \left[36y - 3y^3\right]_0^2 = 6.$$

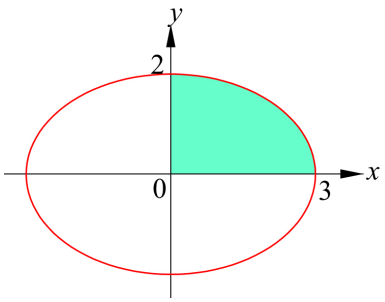


Figure 6.28: The domain $\mathfrak{D} = \{(x, y) \,|\, x \geq 0, y \geq 0, 4x^2 + 9y^2 \leq 36\}$.

Now let us generalize the Fubini's theorem to arbitrary positive integer n that is larger than 1.

Theorem 6.56 Fubini's Theorem

Let n be a positive integer larger than 1, and let k be a positive integer less than n. Given that $\mathbf{I} = \prod_{i=1}^{n}[a_i, b_i]$ is a closed rectangle in $\mathbb{R}^n$, let $\mathbf{I_x} = \prod_{i=1}^{k}[a_i, b_i]$ and $\mathbf{I_y} = \prod_{i=k+1}^{n}[a_i, b_i]$. Assume that $f : \mathbf{I} \to \mathbb{R}$ is a Riemann integrable function such that for each $\mathbf{x}$ in $\mathbf{I_x}$, the function $g_{\mathbf{x}} : \mathbf{I_y} \to \mathbb{R}$ defined by

$$g_{\mathbf{x}}(\mathbf{y}) = f(\mathbf{x}, \mathbf{y}), \quad \mathbf{y} \in \mathbf{I_y}$$

is Riemann integrable. Let $F : \mathbf{I_x} \to \mathbb{R}$ be the function defined as

$$F(\mathbf{x}) = \int_{\mathbf{I_y}} g_{\mathbf{x}}(\mathbf{y})d\mathbf{y} = \int_{\mathbf{I_y}} f(\mathbf{x}, \mathbf{y})d\mathbf{y}.$$

Then we have the followings.

(a) The function $F : \mathbf{I_x} \to \mathbb{R}$ is Riemann integrable.

(b) The integral of $f : \mathbf{I} \to \mathbb{R}$ is equal to the integral of $F : \mathbf{I_x} \to \mathbb{R}$. Namely,

$$\int_{\mathbf{I}} f(\mathbf{x}, \mathbf{y})d\mathbf{x}d\mathbf{y} = \int_{\mathbf{I_x}} \int_{\mathbf{I_y}} f(\mathbf{x}, \mathbf{y})d\mathbf{y}d\mathbf{x}.$$

(c) For each $\mathbf{y}$ in $\mathbf{I_y}$, define the function $h_{\mathbf{y}} : \mathbf{I_x} \to \mathbb{R}$ by

$$h_{\mathbf{y}}(\mathbf{x}) = f(\mathbf{x}, \mathbf{y}), \quad \mathbf{x} \in \mathbf{I_x}.$$

If the function $h_{\mathbf{y}} : \mathbf{I_x} \to \mathbb{R}$ is Riemann integrable for each $\mathbf{y} \in \mathbf{I_y}$, then we can interchange the order of integration. Namely,

$$\int_{\mathbf{I_x}} \int_{\mathbf{I_y}} f(\mathbf{x}, \mathbf{y})d\mathbf{y}d\mathbf{x} = \int_{\mathbf{I_y}} \int_{\mathbf{I_x}} f(\mathbf{x}, \mathbf{y})d\mathbf{x}d\mathbf{y}.$$

The proof is similar to the $n = 2$ case and we leave it to the readers.
A useful case is the following which generalizes Theorem 6.55.

Theorem 6.57

Let $\mathcal{U}$ be a Jordan measurable set in $\mathbb{R}^{n-1}$, and let $g : \mathcal{U} \to \mathbb{R}$ and $h : \mathcal{U} \to \mathbb{R}$ be bounded continuous functions on $\mathcal{U}$ satisfying $g(\mathbf{x}) \leq h(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{U}$. Consider the subset $\mathfrak{D}$ of $\mathbb{R}^n$ defined as

$$\mathfrak{D} = \{(\mathbf{x}, y) \,|\, \mathbf{x} \in \mathcal{U}, g(\mathbf{x}) \leq y \leq h(\mathbf{x})\}.$$

If $f : \mathfrak{D} \to \mathbb{R}$ is a bounded continuous function, then it is Riemann integrable, and

$$\int_{\mathfrak{D}} f(\mathbf{x}, y) d\mathbf{x} dy = \int_{\mathcal{U}} \int_{g(\mathbf{x})}^{h(\mathbf{x})} f(\mathbf{x}, y) dy d\mathbf{x},$$

Let us look at an example.

Example 6.44

Evaluate the integral $\int_{\mathcal{S}} x dx dy dz$, where $\mathcal{S}$ is the solid bounded between the plane $x + y + z = 1$ and the three coordinate planes. Then find the integral $\int_{\mathcal{S}} (x + 5y + 3z) dx dy dz$.

Solution

The solid $\mathcal{S}$ can be expressed as

$$\mathcal{S} = \{(x, y, z) \,|\, (x, y) \in \mathfrak{D}, 0 \leq z \leq 1 - x - y\},$$

where

$$\mathfrak{D} = \{(x, y) \,|\, 0 \leq x \leq 1, 0 \leq y \leq 1 - x\}.$$

Since $\mathfrak{D}$ is a triangle, it is a Jordan measurable set. The function $f(x, y, z) = x$ is continuous.

Hence, we can apply Theorem 6.57.

$$\begin{aligned}\int_{\mathcal{S}} x dxdydz &= \int_{\mathfrak{D}} \left(\int_0^{1-x-y} x dz \right) dxdy \\ &= \int_0^1 \int_0^{1-x} x(1-x-y) dydx \\ &= \int_0^1 x \left[(1-x)y - \frac{y^2}{2} \right]_0^{1-x} dx \\ &= \frac{1}{2} \int_0^1 x(1-x)^2 dx = \frac{1}{2} \int_0^1 x^2(1-x) dx \\ &= \frac{1}{2} \left(\frac{1}{3} - \frac{1}{4} \right) = \frac{1}{24}.\end{aligned}$$

Since the solid $\mathcal{S}$ is symmetric in x, y and z, we have

$$\int_{\mathcal{S}} x dxdydz = \int_{\mathcal{S}} y dxdydz = \int_{\mathcal{S}} z dxdydz.$$

Therefore,

$$\int_{\mathcal{S}} (x + 5y + 3z) dxdydz = 9 \int_{\mathcal{S}} x dxdydz = \frac{3}{8}.$$

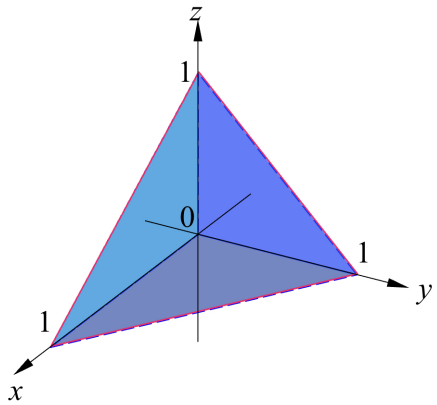


Figure 6.29: The solid $\mathcal{S}$ bounded between the plane $x + y + z = 1$ and the three coordinate planes.

Exercises 6.4

Question 1

Let $\mathbf{I} = [0, 2] \times [0, 2]$, and let $f : \mathbf{I} \to \mathbb{R}$ be the function defined as $f(x, y) = x^7 y^3$. For a positive integer k, let $\mathbf{P}_k$ be the uniformly regular partition of $\mathbf{I}$ into k^2 rectangles. Write down the summation formula for the Darboux upper sum $U(f, \mathbf{P}_k)$. Show that the limit $\lim_{k\to\infty} U(f, \mathbf{P}_k)$ exists and find the limit.

Question 2

Let $\mathfrak{D}$ be the triangle with vertices $(0, 0)$, $(1, 0)$ and $(1, 1)$. Evaluate the integral $\int_{\mathfrak{D}} e^{x^2} dxdy$.

Question 3

Let $\mathfrak{D}$ be the region in the plane bounded between the cuve $y = x^2$ and the line $y = 2x + 3$. Evaluate the integral $\int_{\mathfrak{D}} (x + 2y)dxdy$.

Question 4

Let $\mathfrak{D}$ be the region in the plane bounded between the cuve $y^2 = 4x$ and the line $y = 2x - 4$. Evaluate the integral $\int_{\mathfrak{D}} (x + y)dxdy$.

Question 5

Evaluate the integral $\int_0^1 \int_y^1 \sqrt{9x^2 + 16}\, dxdy$.

Question 6

Evaluate the integral $\int_{\mathcal{S}} xydxdydz$, where $\mathcal{S}$ is the solid bounded between the plane $x + y + z = 4$ and the three coordinate planes. Then find the integral $\int_{\mathcal{S}} (4xy + 5yz + 6xz)dxdydz$.

Question 7

Let $f : [a, b] \to \mathbb{R}$ be a continuous function, and let G be the set

$$G = \{(x, y, 0) \,|\, a \leq x \leq b, 0 \leq y \leq |f(x)|\}$$

in $\mathbb{R}^3$ that lies in the plane $z = 0$. Rotate the set G about the x-axis, we obtain a solid of revolution S, which can be described as

$$S = \left\{(x, y, z) \,|\, a \leq x \leq b, y^2 + z^2 \leq f(x)^2\right\}.$$

Show that the volume of S is

$$\text{vol}\,(S) = \pi \int_a^b f(x)^2 dx.$$

Question 8

Let $\mathfrak{D}_1$ and $\mathfrak{D}_2$ be Jordan measurable sets in $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively. Show that the set $\mathfrak{D} = \mathfrak{D}_1 \times \mathfrak{D}_2$ is a Jordan measurable set in $\mathbb{R}^{m+n}$ and

$$\text{vol}\,(\mathfrak{D}_1 \times \mathfrak{D}_2) = \text{vol}\,(\mathfrak{D}_1) \times \text{vol}\,(\mathfrak{D}_2).$$

6.5 Change of Variables Theorem

Consider the problem of evaluating an integral of the form $\int_{\mathfrak{D}} f(x,y)dxdy$ when $\mathfrak{D}$ is the disc $\mathfrak{D} = \{(x,y) \,|\, x^2+y^2 \le r^2\}$. When $f : \mathfrak{D} \to \mathbb{R}$ is a continuous function, Fubini's theorem says that we can write the integral as

$$\int_{\mathfrak{D}} f(x,y)dxdy = \int_{-r}^{r} \int_{-\sqrt{r^2-x^2}}^{\sqrt{r^2-x^2}} f(x,y)dydx.$$

However, it is usually quite complicated to evaluate this integral due to the square roots. In some sense, we have not fully utilized the circular symmetry of the region of integration $\mathfrak{D}$. For regions that have circular symmetry, it might be easier if we use polar coordinates (r,θ) instead of rectangular coordinates (x,y). The goal of this section is to discuss the change of variables formula for multiple integrals.

For single variable functions, the change of variable formula is usually known as integration by substitution. We have proved the following theorem in volume I.

Theorem 6.58 Integration by Substitution

Let $\psi : [a,b] \to \mathbb{R}$ be a function that satisfies the following conditions:

(i) ψ is continuous and one-to-one on $[a,b]$;

(ii) ψ is continuously differentiable on (a,b);

(iii) $\psi'(x)$ is bounded on (a,b).

If $\psi([a,b]) = [c,d]$, and $f : [c,d] \to \mathbb{R}$ is a bounded function that is continuous on (c,d), then the function $h : [a,b] \to \mathbb{R}$,

$$h(x) = f(\psi(x))|\psi'(x)|$$

is Riemann integrable and

$$\int_c^d f(u)du = \int_a^b f(\psi(x))|\psi'(x)|dx.$$

The function $\psi : [a,b] \to \mathbb{R}$ that satisfies all the three conditions (i)–(iii) in Theorem 6.58 defines a *smooth* change of variables $u = \psi(x)$ from x to u.

Definition 6.18 Smooth Change of Variables

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$. A mapping $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^n$ from $\mathcal{O}$ to $\mathbb{R}^n$ is called a smooth change of variables provided that it satisfies the following conditions.

(i) $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^n$ is one-to-one.

(ii) $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^n$ is continuously differentiable.

(iii) For each $\mathbf{x} \in \mathcal{O}$, the derivative matrix $\mathbf{D\Psi}(\mathbf{x})$ is invertible.

Remark 6.8

If the mapping $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^n$ is continuously differentiable, and the derivative matrix $\mathbf{D\Psi}(\mathbf{x})$ is invertible for each $\mathbf{x} \in \mathcal{O}$, the inverse function theorem implies that the mapping $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^n$ is locally one-to-one. However, it might not be *globally one-to-one*. For $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^n$ to be a smooth change of variables, we need to impose the additional condition that it is globally one-to-one.

Example 6.45

Let $\mathbf{x}_0$ be a point in $\mathbb{R}^n$. The mapping $\mathbf{\Psi} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{\Psi}(\mathbf{x}) = \mathbf{x} + \mathbf{x}_0$ is a smooth change of variables. It is one-to-one, continuously differentiable, and the derivative matrix is $\mathbf{D\Psi}(\mathbf{x}) = I_n$, which is invertible.

Example 6.46

Let $\mathbf{x}_0$ and $\mathbf{y}_0$ be points in $\mathbb{R}^n$, and let A be an invertible $n \times n$ matrix. The mapping $\mathbf{\Psi} : \mathbb{R}^n \to \mathbb{R}^n$ defined by

$$\mathbf{\Psi}(\mathbf{x}) = \mathbf{y}_0 + A(\mathbf{x} - \mathbf{x}_0)$$

is a one-to-one continuously differentiable mapping. Its derivative matrix is $\mathbf{D\Psi}(\mathbf{x}) = A$, which is invertible for all $\mathbf{x}$ in $\mathbb{R}^n$. This shows that $\mathbf{\Psi}$ is a smooth change of variables.

The mapping $\mathbf{\Psi} : \mathbb{R}^n \to \mathbb{R}^n$ in Example 6.46 is a composition of translations

and an invertible linear transformation.

Example 6.47

Let $\mathcal{O} = \{(x, y) \,|\, x > 0, y > 0\}$, and let $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^2$ be the mapping defined as

$$\mathbf{\Psi}(x, y) = (x^2 - y^2, 2xy).$$

Show that $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^2$ is a smooth change of variables.

Solution

First we show that $\mathbf{\Psi}$ is one-to-one. If $\mathbf{\Psi}(x_1, y_1) = \mathbf{\Psi}(x_2, y_2)$, then

$$x_1^2 - y_1^2 = x_2^2 - y_2^2, \quad 2x_1y_1 = 2x_2y_2.$$

Let $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$. Then we find that

$$z_1^2 = x_1^2 - y_1^2 + 2ix_1y_1 = x_2^2 - y_2^2 + 2ix_2y_2 = z_2^2.$$

Hence, we must have $z_2 = \pm z_1$. Restricted to $\mathcal{O}$, $x_1, x_2, y_1, y_2 > 0$. Hence, we must have $z_1 = z_2$, or equivalently, $(x_1, y_1) = (x_2, y_2)$. This shows that $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^2$ is one-to-one. Now

$$\mathbf{D\Psi}(x, y) = \begin{bmatrix} 2x & -2y \\ 2y & 2x \end{bmatrix}$$

is continuous, and $\det \mathbf{D\Psi}(x, y) = 4x^2 + 4y^2 \neq 0$ for all $(x, y) \in \mathcal{O}$. This proves that $\mathbf{\Psi}$ is continuously differentiable and the derivative matrix $\mathbf{D\Psi}(x, y)$ is invertible for all $(x, y) \in \mathcal{O}$. Hence, $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^2$ is a smooth change of variables.

In this section, we will state the change of variables theorem, and give some discusssions about why this theorem holds. We will also look at examples of how this theorem is applied, especially for polar and spherical coordinates. The proof of the theorem is quite technical and will be given in next section.

Theorem 6.59 The Change of Variables Theorem

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^n$ be a smooth change of variables. If $\mathfrak{D}$ is a Jordan measurable set such that its closure $\overline{\mathfrak{D}}$ is contained in $\mathcal{O}$, then $\mathbf{\Psi}(\mathfrak{D})$ is also Jordan measurable. If $f : \mathbf{\Psi}(\mathfrak{D}) \to \mathbb{R}$ is a bounded continuous function, then the function $g : \mathfrak{D} \to \mathbb{R}$ defined as

$$g(\mathbf{x}) = f\left(\mathbf{\Psi}(\mathbf{x})\right) \left|\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})\right|$$

is Riemann integrable, and

$$\int_{\mathbf{\Psi}(\mathfrak{D})} f(\mathbf{x}) d\mathbf{x} = \int_{\mathfrak{D}} g(\mathbf{x}) d\mathbf{x} = \int_{\mathfrak{D}} f\left(\mathbf{\Psi}(\mathbf{x})\right) \left|\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})\right| d\mathbf{x}. \tag{6.5}$$

Notice that the two vertical lines on $\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})$ in (6.5) means the absolute value, not the determinant.

Remark 6.9 Jacobian

For a mapping $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^n$ from a subset of $\mathbb{R}^n$ to $\mathbb{R}^n$, the derivative matrix $\mathbf{D}\mathbf{\Psi}(\mathbf{x})$ is also called the *Jacobian matrix* of the mapping $\mathbf{\Psi}$. The determinant of the Jacobian matrix is denoted by

$$\frac{\partial(\Psi_1, \ldots, \Psi_n)}{\partial(x_1, \ldots, x_n)}.$$

It is known as the *Jacobian determinant*, or simply as *Jacobian*. In practice, we will often denote a change of variables $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^n$ by $\mathbf{u} = \mathbf{\Psi}(\mathbf{x})$. Then the Jacobian can be written as

$$\frac{\partial(u_1, \ldots, u_n)}{\partial(x_1, \ldots, x_n)}.$$

Using this notation, the change of variables formula (6.5) reads as

$$\begin{aligned} &\int_{\mathbf{\Psi}(\mathfrak{D})} f(u_1, \ldots, u_n) du_1 \cdots du_n \\ &= \int_{\mathfrak{D}} f(u_1(\mathbf{x}), \ldots, u_n(\mathbf{x})) \left|\frac{\partial(u_1, \ldots, u_n)}{\partial(x_1, \ldots, x_n)}\right| dx_1 \ldots dx_n. \end{aligned}$$

6.5.1 Translations and Linear Transformations

In the single variable case, a translation is a map $T : \mathbb{R} \to \mathbb{R}$, $T(x) = x + c$. If $f : [a, b] \to \mathbb{R}$ is a Riemann integrable function, then

$$\int_a^b f(x)dx = \int_{a-c}^{b-c} f(x + c)dx.$$

For $n \geq 2$, we have the following theorem, which is a stronger version of Theorem 6.59.

Theorem 6.60

Let $\mathbf{x}_0$ be a fixed point in $\mathbb{R}^n$, and let $\mathbf{\Psi} : \mathbb{R}^n \to \mathbb{R}^n$ be the translation $\mathbf{\Psi}(\mathbf{x}) = \mathbf{x} + \mathbf{x}_0$. If $\mathfrak{D}$ is a Jordan measurable subset of $\mathbb{R}^n$, then $\mathbf{\Psi}(\mathfrak{D})$ is Jordan measurable. If $f : \mathbf{\Psi}(\mathfrak{D}) \to \mathbb{R}$ is a Riemann integrable function, then $g = (f \circ \mathbf{\Psi}) : \mathfrak{D} \to \mathbb{R}$ is also Riemann integrable, and

$$\int_{\mathfrak{D}+\mathbf{x}_0} f(\mathbf{x})d\mathbf{x} = \int_{\mathbf{\Psi}(\mathfrak{D})} f(\mathbf{x})d\mathbf{x} = \int_{\mathfrak{D}} g(\mathbf{x})d\mathbf{x} = \int_{\mathfrak{D}} f(\mathbf{x} + \mathbf{x}_0)d\mathbf{x}. \quad (6.6)$$

Proof

Obviously, translation maps a rectangle to a rectangle with the same volume. Hence, it maps sets that have Jordan content zero to sets that have Jordan content zero. It is also obvious that $\mathbf{\Psi}$ maps the boundary of $\mathfrak{D}$ to the boundary of $\mathbf{\Psi}(\mathfrak{D})$. This shows that $\mathbf{\Psi}(\mathfrak{D})$ is Jordan measurable.
If $\mathbf{I}$ is a closed rectangle that contains $\mathfrak{D}$, then $\mathbf{I}' = \mathbf{I} + \mathbf{x}_0$ is a closed rectangle that contains $\mathbf{\Psi}(\mathfrak{D})$. Let $\check{f} : \mathbf{I}' \to \mathbb{R}$ and $\check{g} : \mathbf{I} \to \mathbb{R}$ be the zero extensions of $f : \mathbf{\Psi}(\mathfrak{D}) \to \mathbb{R}$ and $g = (f \circ \mathbf{\Psi}) : \mathfrak{D} \to \mathbb{R}$ respectively. Then $\check{g} = \check{f} \circ \mathbf{\Psi}$. Since $f : \mathbf{\Psi}(\mathfrak{D}) \to \mathbb{R}$ is a Riemann integrable,

$$\underline{\int_{\mathbf{I}'}} \check{f} = \overline{\int_{\mathbf{I}'}} \check{f} = \int_{\mathbf{I}'} \check{f} = \int_{\mathbf{\Psi}(\mathfrak{D})} f.$$

Given a partition $\mathbf{P}' = (P'_1, \ldots, P'_n)$ of $\mathbf{I}'$, let $\mathbf{P} = (P_1, \ldots, P_n)$ be the partition of $\mathbf{I}$ induced by the translation $\mathbf{\Psi}$. Namely, $\mathbf{P}$ is a partition such that the rectangle $\mathbf{J}$ is in $\mathcal{J}_{\mathbf{P}}$ if and only if $\mathbf{J}' = \mathbf{\Psi}(\mathbf{J}) = \mathbf{J} + \mathbf{x}_0$ is in $\mathbf{P}'$.

Then

$$\begin{aligned} m_{\mathbf{J}}(\check{g}) &= \inf\{g(\mathbf{x}) \,|\, \mathbf{x} \in \mathbf{J}\} = \inf\{f(\mathbf{x}+\mathbf{x}_0) \,|\, \mathbf{x} \in \mathbf{J}\} \\ &= \inf\{f(\mathbf{x}) \,|\, \mathbf{x}' \in \mathbf{J}+\mathbf{x}_0\} = m_{\mathbf{J}'}(\check{f}). \end{aligned}$$

It follows that

$$L(\check{g}, \mathbf{P}) = \sum_{\mathbf{J}\in\mathcal{J}_{\mathbf{P}}} m_{\mathbf{J}}(\check{g}) \operatorname{vol}(\mathbf{J}) = \sum_{\mathbf{J}'\in\mathcal{J}_{\mathbf{P}'}} m_{\mathbf{J}'}(\check{f}) \operatorname{vol}(\mathbf{J}') = L(\check{f}, \mathbf{P}').$$

Similarly, we have

$$U(\check{g}, \mathbf{P}) = U(\check{f}, \mathbf{P}').$$

Thus, the sets $S_L(\check{g})$ and $S_L(\check{f})$ of lower sums of $\check{g}$ and $\check{f}$ are the same, and the sets $S_U(\check{g})$ and $S_U(\check{f})$ of upper sums of $\check{g}$ and $\check{f}$ are also the same. These imply that

$$\begin{aligned} \underline{\int_{\mathbf{I}}} \check{g} &= \sup S_L(\check{g}) = \sup S_L(\check{f}) = \underline{\int_{\mathbf{I}'}} \check{f} = \int_{\mathbf{\Psi}(\mathfrak{D})} f, \\ \overline{\int_{\mathbf{I}}} \check{g} &= \inf S_U(\check{g}) = \inf S_U(\check{f}) = \overline{\int_{\mathbf{I}'}} \check{f} = \int_{\mathbf{\Psi}(\mathfrak{D})} f. \end{aligned}$$

Hence, $g : \mathfrak{D} \to \mathbf{I}$ is Riemann integrable and

$$\int_{\mathfrak{D}} g = \int_{\mathbf{I}} \check{g} = \int_{\mathbf{\Psi}(\mathfrak{D})} f.$$

Remark 6.10

It is easy to check that for the translation $\mathbf{\Psi} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{\Psi}(\mathbf{x}) = \mathbf{x} + \mathbf{x}_0$, the change of variables formula (6.6) is precisely the formula (6.5), since $\mathbf{D}\mathbf{\Psi}(\mathbf{x}) = I_n$ in this case.

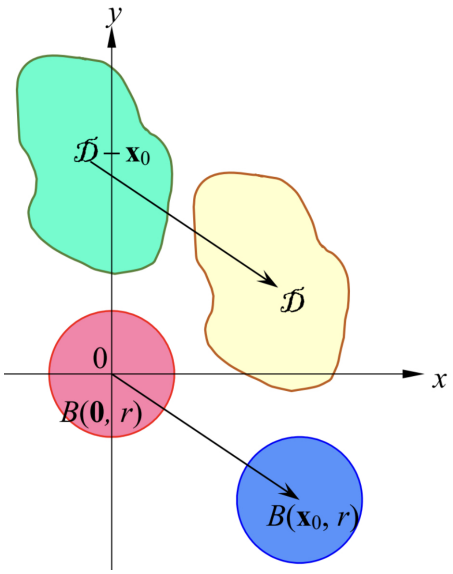


Figure 6.30: A translation in the plane.

Example 6.48

Let

$$\mathfrak{D} = \left\{(x, y) \,|\, (x-2)^2 + (y+3)^2 \leq 16\right\}.$$

Evaluate the integral $\displaystyle\int_{\mathfrak{D}} (4x + y)dxdy$.

Solution

Make the change of variables $u = x - 2$, $v = y + 3$, which is a translation. Then $x = u + 2$, $y = v - 3$, and we have

$$\begin{aligned}\int_{\mathfrak{D}} (4x + y)dxdy &= \int_{u^2+v^2\leq 16} (4u + 8 + v - 3)dudv \\ &= \int_{u^2+v^2\leq 16} (4u + v + 5)dudv.\end{aligned}$$

Since the disc $\mathcal{B} = \{(u, v) \,|\, u^2 + v^2 \leq 16\}$ is invariant when we change u to $-u$, or change v to $-v$, the integrals

$$\int_{u^2+v^2\le 16} u\,du\,dv \quad \text{and} \quad \int_{u^2+v^2\le 16} v\,du\,dv$$

are equal to 0. Therefore,

$$\int_{\mathfrak{D}} (4x+y)\,dx\,dy = 5\int_{u^2+v^2\le 16} du\,dv.$$

In single variable analysis, we have shown that the area of a disc of radius r is πr^2. Hence,

$$\int_{\mathfrak{D}} (4x+y)\,dx\,dy = 5 \times \text{area}\,(\mathcal{B}) = 5 \times 16\pi = 80\pi.$$

Now we consider a linear transformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{T}(\mathbf{x}) = A\mathbf{x}$ defined by an invertible matrix A. Since $\mathbf{DT}(\mathbf{x}) = A$, the change of variables theorem says that for any function $f : \mathfrak{D} \to \mathbb{R}$ that is bounded and continuous on $\mathfrak{D}$,

$$\int_{\mathbf{T}(\mathfrak{D})} f(\mathbf{x})\,d\mathbf{x} = \int_{\mathfrak{D}} f(\mathbf{T}(\mathbf{x}))|\det A|\,d\mathbf{x} = |\det A| \int_{\mathfrak{D}} f(\mathbf{T}(\mathbf{x}))\,d\mathbf{x}. \tag{6.7}$$

In the special case where f is a constant function, we have

$$\text{vol}\,(\mathbf{T}(\mathfrak{D})) = |\det A|\,\text{vol}\,(\mathfrak{D}).$$

A very crucial fact to the proof of the change of variables theorem is a special case of this formula when $\mathfrak{D}$ is a rectangle.

Theorem 6.61

Let $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$, and let $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{T}(\mathbf{x}) = A\mathbf{x}$ be an invertible linear transformation. Then

$$\text{vol}\,(\mathbf{T}(\mathbf{I})) = |\det A|\,\text{vol}\,(\mathbf{I}). \tag{6.8}$$

Linear transformations map linear objects to linear objects. However, the image of a rectangle under a linear transformation is not necessary a rectangle, but is always a parallelepiped. If $\mathbf{I}$ is the closed rectangle $\prod_{i=1}^{n} [a_i, b_i]$, a point $\mathbf{x}$ in $\mathbf{I}$

can be written as

$$\mathbf{x} = \mathbf{a} + t_1(b_1 - a_1)\mathbf{e}_1 + \cdots + t_n(b_n - a_n)\mathbf{e}_n,$$

where $\mathbf{a} = (a_1, \ldots, a_n)$, and $\mathbf{t} = (t_1, \ldots, t_n) \in [0,1]^n$. Hence, we say that the rectangle $\mathbf{I}$ is a parallelepiped based at the point $\mathbf{a}$ and spanned by the n-linearly independent vectors $\mathbf{v}_i = (b_i - a_i)\mathbf{e}_i$, $1 \leq i \leq n$.

Definition 6.19 Parallelepipeds

A (closed) parallelepiped in $\mathbb{R}^n$ is a solid $\mathscr{P}$ in $\mathbb{R}^n$ based at a point $\mathbf{a}$ and spanned by n-linearly independent vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$. It can be described as

$$\mathscr{P} = \{\mathbf{a} + t_1\mathbf{v}_1 + \cdots + t_n\mathbf{v}_n \,|\, \mathbf{t} = (t_1, \ldots, t_n) \in [0,1]^n\}.$$

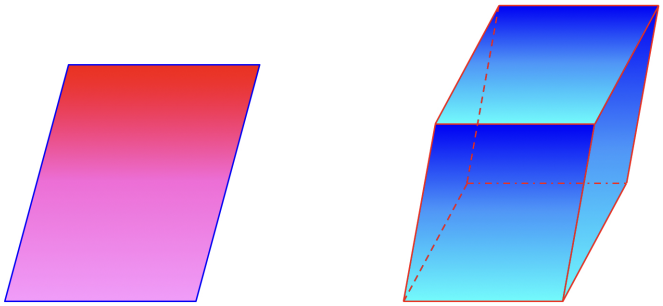


Figure 6.31: Parallelepipeds in $\mathbb{R}^2$ and $\mathbb{R}^3$.

The boundary of a parallelepiped is a union of $2n$ bounded subsets, each of them is contained in a hyperplane. Thus, the boundary of a parallelepiped has Jordan content zero. Therefore, a parallelepiped is a Jordan measurable set.

If $\mathscr{P}$ be a parallelepiped in $\mathbb{R}^n$ based at the point $\mathbf{a}$ and spanned by the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$, and $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$ is an invertible linear transformation, then $\mathbf{T}(\mathscr{P})$ is the parallelepiped in $\mathbb{R}^n$ based at the point $\mathbf{T}(\mathbf{a})$ and spanned by the vectors $\mathbf{T}(\mathbf{v}_1), \ldots, \mathbf{T}(\mathbf{v}_n)$.

The cube $[0,1]^n$ is called the standard unit cube and it is often denoted by Q_n. If $\mathscr{P}$ is a parallelepiped in $\mathbb{R}^n$ based at the point $\mathbf{a}$ and spanned by the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$, then $\mathscr{P} = \mathbf{\Psi}(Q_n)$, where $\mathbf{\Psi} : \mathbb{R}^n \to \mathbb{R}^n$ is the mapping

$$\mathbf{\Psi}(\mathbf{x}) = A\mathbf{x} + \mathbf{a}, \qquad A = \left[\mathbf{v}_1 \,\middle|\, \cdots \,\middle|\, \mathbf{v}_n\right],$$

which is a composition of an invertible linear transformation and a translation. Theorem 6.61 says that

$$\text{vol}\,(\mathscr{P}) = |\det A|, \tag{6.9}$$

where A is the matrix whose column vectors are $\mathbf{v}_1, \ldots, \mathbf{v}_n$. For example, for a parallelogram in $\mathbb{R}^2$ which is spanned by the vectors

$$\mathbf{v}_1 = \begin{bmatrix} a_1 \\ b_1 \end{bmatrix} \quad \text{and} \quad \mathbf{v}_2 = \begin{bmatrix} a_2 \\ b_2 \end{bmatrix},$$

the area of the parallogram is

$$\det \begin{bmatrix} a_1 & a_2 \\ b_1 & b_2 \end{bmatrix}.$$

For a parallelepiped in $\mathbb{R}^3$ which is spanned by the vectors

$$\mathbf{v}_1 = \begin{bmatrix} a_1 \\ b_1 \\ c_1 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} a_2 \\ b_2 \\ c_2 \end{bmatrix} \quad \text{and} \quad \mathbf{v}_3 = \begin{bmatrix} a_3 \\ b_3 \\ c_3 \end{bmatrix},$$

the volume of the parallelepiped is

$$\det \begin{bmatrix} a_1 & a_1 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{bmatrix}.$$

These formulas have been derived in an elementary course. For general n, we will prove (6.9) in Appendix B using geometric arguments. This will then imply Theorem 6.61.

From the theory of linear algebra, we know that an invertible matrix is a product of elementary matrices. Hence, an invertible linear transformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$ can be written as

$$\mathbf{T} = \mathbf{T}_m \circ \cdots \circ \mathbf{T}_2 \circ \mathbf{T}_1,$$

where $\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_m$ is one of the three types of elementary transformations, corresponding to the three types of elementary matrices.

I. When E is the elementary matrix obtained from the identity matrix I_n by interchanging two distinct rows i and j, the linear tranformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{T}(\mathbf{x}) = E\mathbf{x}$ interchanges x_i and x_j, and fixes the other variables. In this case, $\det E = -1$ and $|\det E| = 1$.

II. When E is the elementary matrix obtained from the identity matrix I_n by multiplying row i by a nonzero constant c, the linear transformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{T}(\mathbf{x}) = E\mathbf{x}$ maps the point $\mathbf{x} = (x_1, \ldots, x_n)$ to

$$\mathbf{T}(\mathbf{x}) = (x_1, \ldots, x_{i-1}, cx_i, x_{i+1}, \ldots, x_n).$$

In this case, $\det E = c$, and $|\det E| = |c|$.

III. When E is the elementary matrix obtained from the identity matrix I_n by adding a constant c times of row j to another row i, the linear transformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{T}(\mathbf{x}) = E\mathbf{x}$ maps the point $\mathbf{x} = (x_1, \ldots, x_n)$ to

$$\mathbf{T}(\mathbf{x}) = (x_1, \ldots, x_{i-1}, x_i + cx_j, x_{i+1}, \ldots, x_n).$$

In this case, $\det E = 1$, and $|\det E| = 1$.

Since each of the elementary transformations involves changes in at most two variables, it is sufficient to consider these transformations when $n = 2$.

Example 6.49

Let $\mathbf{T} : \mathbb{R}^2 \to \mathbb{R}^2$ be the linear transformation

$$\mathbf{T}(x, y) = (y, x).$$

The matrix E corresponding to this transformation is

$$E = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Under this transformation, the rectangle $\mathbf{I} = [a, b] \times [c, d]$ is mapped to the rectangle $\mathbf{I}' = [c, d] \times [a, b]$. It is easy to see that

$$\text{vol}\,(\mathbf{I}') = \text{vol}\,(\mathbf{I}).$$

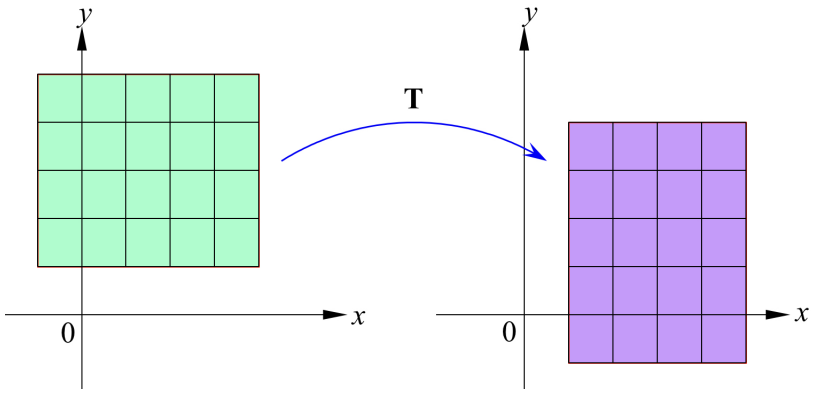


Figure 6.32: The linear transformation $\mathbf{T}(x, y) = (y, x)$.

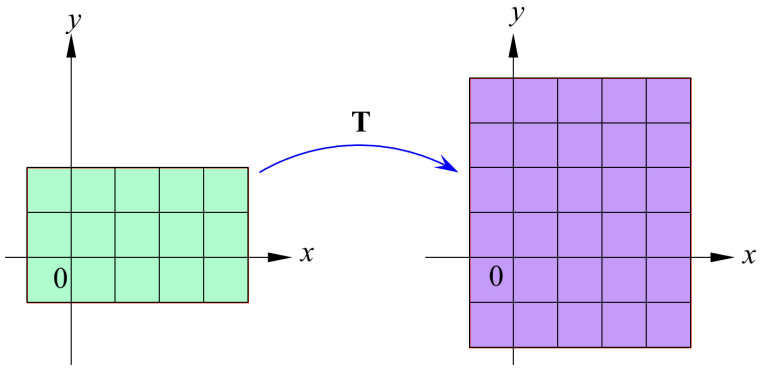


Figure 6.33: The linear transformation $\mathbf{T}(x, y) = (x, 2y)$.

Example 6.50

Let $\mathbf{T} : \mathbb{R}^2 \to \mathbb{R}^2$ be the linear transformation

$$\mathbf{T}(x, y) = (x, ky), \qquad k \neq 0.$$

The matrix E corresponding to this transformation is

$$E = \begin{bmatrix} 1 & 0 \\ 0 & k \end{bmatrix}.$$

Under this transformation, the rectangle $\mathbf{I} = [a, b] \times [c, d]$ is mapped to the rectangle $\mathbf{I}' = [a, b] \times [kc, kd]$ if $k > 0$; and to the rectangle $\mathbf{I}' = [a, b] \times [kd, kc]$ if $k < 0$. In any case, we find that

$$\text{vol}\,(\mathbf{I}') = |k|\,\text{vol}\,(\mathbf{I}).$$

Example 6.51

Let $\mathbf{T} : \mathbb{R}^2 \to \mathbb{R}^2$ be the linear transformation

$$\mathbf{T}(x, y) = (x + ky, y).$$

The matrix E corresponding to this transformation is

$$E = \begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix}.$$

Under this transformation, the rectangle $\mathbf{I} = [a, b] \times [c, d]$ is mapped to the parallepiped $\mathscr{P}$ with vertices $(a + kc, c)$, $(a + kd, d)$, $(b + kc, c)$ and $(b + kd, d)$. Using elementary geometric argument, one can show that

$$\text{vol}\,(\mathscr{P}) = \text{vol}\,(\mathbf{I}).$$

Combining Example 6.49, Example 6.50 and Example 6.51, we conclude that (6.8) holds when $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$ is an elementary transformation.

The type II elementary transformations maps rectangles to rectangles, so do their compositions. Therefore, (6.8) also holds if the linear transformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$ is a composition of type II elementary transformations. This gives the following.

Theorem 6.62

Let $\mathbf{x}_0 = (u_1, \ldots, u_n)$ be a fixed point in $\mathbb{R}^n$, and let $\mathbf{\Psi} : \mathbb{R}^n \to \mathbb{R}^n$ be the mapping

$$\Psi_i(\mathbf{x}) = \alpha_i x_i + u_i.$$

Equivalently, $\mathbf{\Psi}(\mathbf{x}) = A\mathbf{x} + \mathbf{x}_0$, where A is a diagonal matrix with diagonal entries $\alpha_1, \ldots, \alpha_n$. If $\mathfrak{D}$ is a Jordan measurable subset of $\mathbb{R}^n$, then $\mathbf{\Psi}(\mathfrak{D})$ is Jordan measurable. If $f : \mathbf{\Psi}(\mathfrak{D}) \to \mathbb{R}$ is a Riemann integrable function, then $h = (f \circ \mathbf{\Psi}) : \mathfrak{D} \to \mathbb{R}$ is also Riemann integrable, and

$$\int_{\mathbf{\Psi}(\mathfrak{D})} f(\mathbf{x})d\mathbf{x} = |\det A| \int_{\mathfrak{D}} h(\mathbf{x})d\mathbf{x} = |\det A| \int_{\mathfrak{D}} f(A\mathbf{x} + \mathbf{x}_0)d\mathbf{x}. \quad (6.10)$$

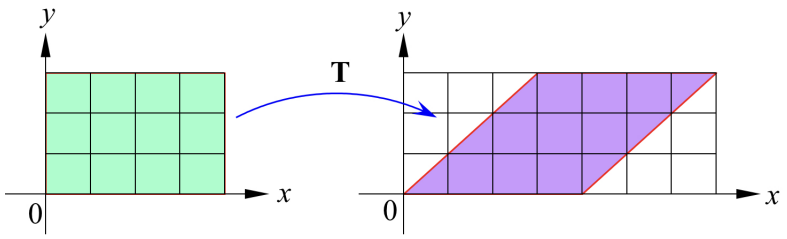


Figure 6.34: The linear transformation $\mathbf{T}(x, y) = (x + y, y)$.

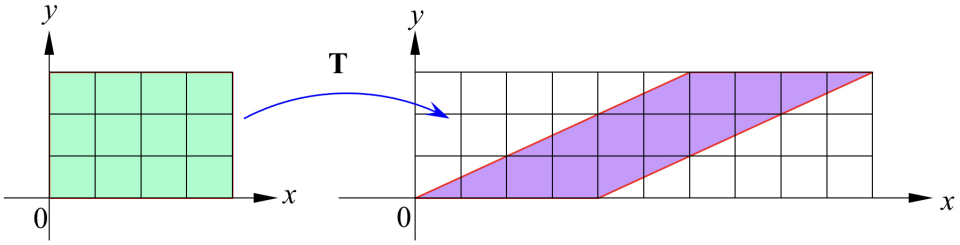


Figure 6.35: The linear transformation $\mathbf{T}(x, y) = (x + 2y, y)$.

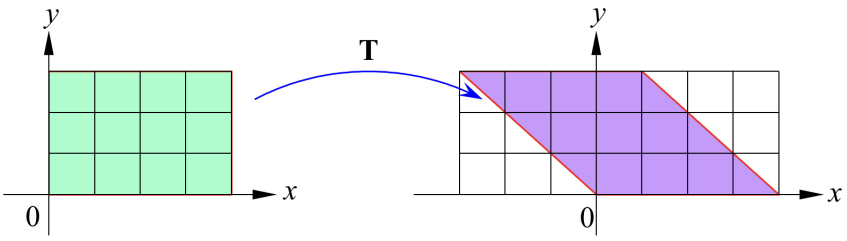


Figure 6.36: The linear transformation $\mathbf{T}(x, y) = (x - y, y)$.

Notice that $\det A = \alpha_1 \cdots \alpha_n$. If $\mathbf{y} = \mathbf{\Psi}(\mathbf{x})$, then

$$y_i = \alpha_i x_i + u_i, \qquad 1 \leq i \leq n.$$

The proof of Theorem 6.62 is similar to the proof Theorem 6.60, by establishing one-to-one correspondence between the partitions, and using the fact that for any rectangles $\mathbf{J}$,

$$\text{vol}\,(\mathbf{\Psi}(\mathbf{J})) = |\alpha_1 \cdots \alpha_n|\,\text{vol}\,(\mathbf{J}).$$

Example 6.52

Find the area of the ellipse

$$\mathcal{E} = \left\{(x, y) \,|\, 4(x+1)^2 + 9(y-5)^2 \leq 49\right\}.$$

Solution

Make a change of variables $u = 2(x+1)$ and $v = 3(y-5)$. The Jacobian is

$$\frac{\partial(u,v)}{\partial(x,y)} = \det \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} = 6,$$

and so

$$\frac{\partial(x,y)}{\partial(u,v)} = \frac{1}{6}.$$

Therefore,

$$\begin{aligned} \text{area}\,(\mathcal{E}) &= \int_{4(x+1)^2+9(y-5)^2 \leq 49} dxdy = \int_{u^2+v^2 \leq 49} \left|\frac{\partial(x,y)}{\partial(u,v)}\right| dudv \\ &= \frac{1}{6}\int_{u^2+v^2 \leq 49} dudv = \frac{49}{6}\pi. \end{aligned}$$

Finally, let us consider an example of applying a general linear transformation.

Example 6.53

Evaluate the integral

$$\int_{\mathfrak{D}} \frac{2x+3y+3}{2x-3y+8} dxdy,$$

where

$$\mathfrak{D} = \{(x, y) \,|\, 2|x| + 3|y| \leq 6\}.$$

Solution

Notice that for any $(x, y) \in \mathfrak{D}$,

$$|2x - 3y + 8| \geq 8 - 2|x| - 3|y| \geq 2.$$

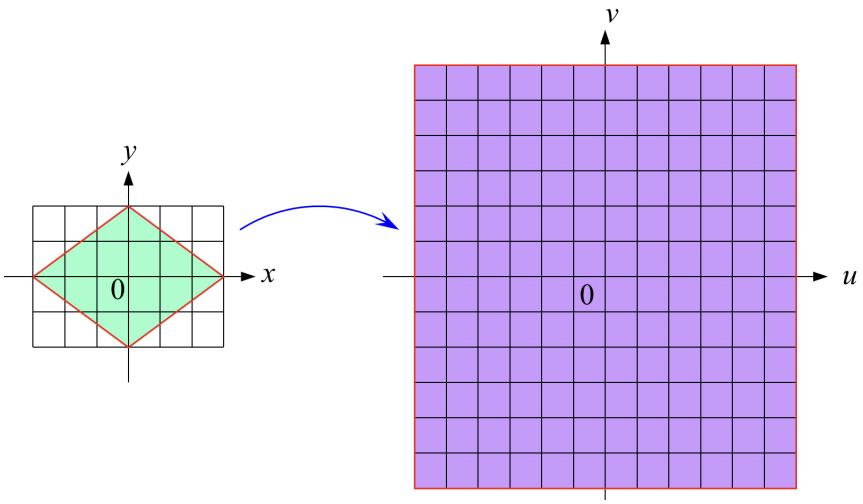


Figure 6.37: The transformation $u = 2x - 3y$ and $v = 2x + 3y$.

Hence, the function

$$h(x, y) = \frac{2x + 3y + 3}{2x - 3y + 8}$$

is continuous on $\mathfrak{D}$. The region $\mathfrak{D}$ is enclosed by the 4 lines $2x + 3y = 6$, $2x + 3y = -6$, $2x - 3y = 6$ and $2x - 3y = -6$. This prompts us to define a change of variables by $u = 2x - 3y$ and $v = 2x + 3y$. This is a linear transformation with Jacobian

$$\frac{\partial(u, v)}{\partial(x, y)} = \det \begin{bmatrix} 2 & -3 \\ 2 & 3 \end{bmatrix} = 12.$$

Therefore,

$$\frac{\partial(x, y)}{\partial(u, v)} = \frac{1}{12}.$$

The region $\mathfrak{D}$ in the (x, y)-plane is mapped to the rectangle

$$\mathcal{R} = \{(u, v) \,|\, -6 \leq u \leq 6, -6 \leq v \leq 6\}$$

in the (u, v)-plane.

Thus,

$$\begin{aligned}\int_{\mathfrak{D}} \frac{2x+3y+3}{2x-3y+8}dxdy &= \int_{\mathcal{R}} \frac{v}{u+8}\left|\frac{\partial(x,y)}{\partial(u,v)}\right| dudv \\ &= \frac{1}{12}\int_{-6}^{6}\int_{-6}^{6} \frac{v+3}{u+8}dudv \\ &= \frac{1}{12}\left[\frac{v^2}{2}+3v\right]_{-6}^{6}[\ln(u+8)]_{-6}^{6} \\ &= 3\ln 7.\end{aligned}$$

6.5.2 Polar Coordinates

Given a point (x, y) in the plane $\mathbb{R}^2$, if r is a nonnegative number and θ is a real number such that

$$x = r\cos\theta, \quad y = r\sin\theta,$$

then (r, θ) are called the polar coordinates of the point (x, y). Notice that

$$r = \sqrt{x^2+y^2}.$$

Restricted to

$$V = \{(r,\theta)\,|\,r > 0, 0 \le \theta < 2\pi\},$$

the map $\mathbf{\Phi} : V \to \mathbb{R}^2$,

$$\mathbf{\Phi}(r,\theta) = (r\cos\theta, r\sin\theta)$$

is one-to-one, and its range is $\mathbb{R}^2 \setminus \{(0,0)\}$. However, the inverse of $\mathbf{\Phi}$ fails to be continuous. We can extend the map $\mathbf{\Phi}$ to $\mathbb{R}^2$ continuously. Namely, given $(r,\theta) \in \mathbb{R}^2$, let $(x, y) = \mathbf{\Phi}(r,\theta)$, where

$$x = r\cos\theta, \quad y = r\sin\theta.$$

Then $\mathbf{\Phi} : \mathbb{R}^2 \to \mathbb{R}^2$ is continuously differentiable, but it fails to be one-to-one. Nevertheless, for any real number α, the map is continuous and one-to-one on the open set

$$\mathcal{O}_\alpha = \{(r,\theta)\,|\,r > 0, \alpha < \theta < \alpha + 2\pi\}.$$

The derivative matrix of $\mathbf{\Phi} : \mathbb{R}^2 \to \mathbb{R}^2$ is given by

$$\mathbf{D\Phi}(r,\theta) = \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix}.$$

Since

$$\det \mathbf{D\Phi}(r,\theta) = r\cos^2\theta + r\sin^2\theta = r,$$

we find that for any $(r,\theta) \in \mathcal{O}_\alpha$, $\mathbf{D\Phi}(r,\theta)$ is invertible. Hence, $\mathbf{\Phi} : \mathcal{O}_\alpha \to \mathbb{R}^2$ is a smooth change of variables.

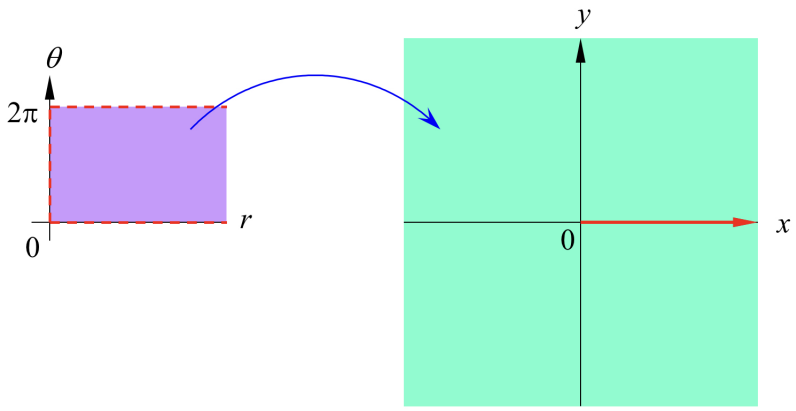


Figure 6.38: The mapping $\mathbf{\Phi} : \mathcal{O} \to \mathbb{R}^2$, $\mathbf{\Phi}(r,\theta) = (r\cos\theta, r\sin\theta)$ maps $\mathcal{O} = \{(r,\theta) \,|\, r > 0, 0 < \theta < 2\pi\}$ to $\mathbb{R}^2 \setminus L$, where L is the positive x-axis.

Let us consider the special case where $\alpha = 0$. In this case, let $\mathcal{O} = \mathcal{O}_0$. The map $\mathbf{\Phi} : \mathcal{O} \to \mathbb{R}^2$,

$$\mathbf{\Phi}(r,\theta) = (r\cos\theta, r\sin\theta)$$

is a smooth change of variables from polar coordinates to rectangular coordinates. Under this change of variables,

$$\mathbf{\Phi}(\mathcal{O}) = \mathbb{R}^2 \setminus \{(x,0) \,|\, x \geq 0\}$$

is an open set in $\mathbb{R}^2$. If $\mathcal{D}$ is the open rectangle $(r_1, r_2) \times (\theta_1, \theta_2)$, with

$$0 < r_1 < r_2 \quad \text{and} \quad 0 < \theta_1 < \theta_2 < 2\pi,$$

then $\mathbf{\Phi}(\mathcal{D})$ is the open set bounded between the two circles $x_1^2 + y_1^2 = r_1^2$ and $x_2^2 + y_2^2 = r_2^2$, and the two rays $y = x\tan\theta_1$, $x \geq 0$ and $y = x\tan\theta_2$, $x \geq 0$.

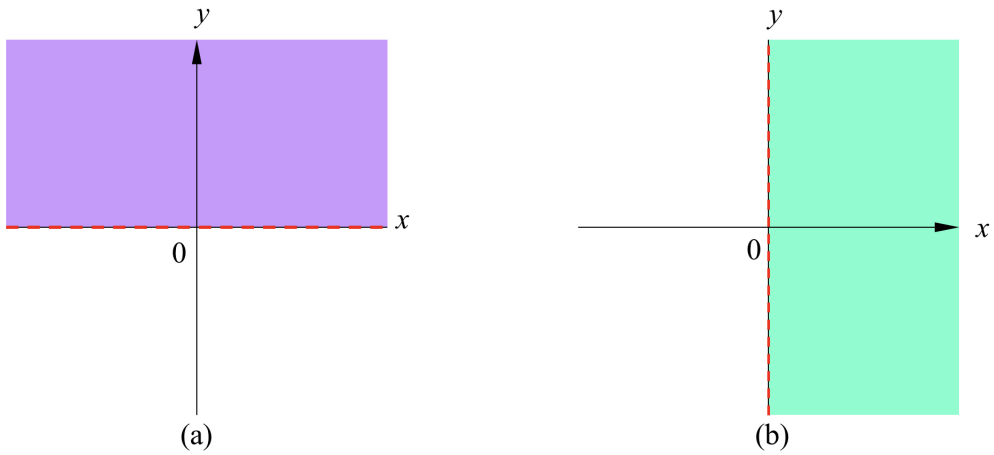


Figure 6.39: The region $\mathfrak{D} = \{(r\cos\theta, r\sin\theta) \,|\, r_1 < r < r_2, \theta_1 < \theta < \theta_2\}$ in the (x, y)-plane. (a) $r_1 = 0$, $r_2 = \infty$, $\theta_1 = 0$, $\theta_2 = \pi$. (b) $r_1 = 0$, $r_2 = \infty$, $\theta_1 = -\frac{\pi}{2}$, $\theta_2 = \frac{\pi}{2}$.

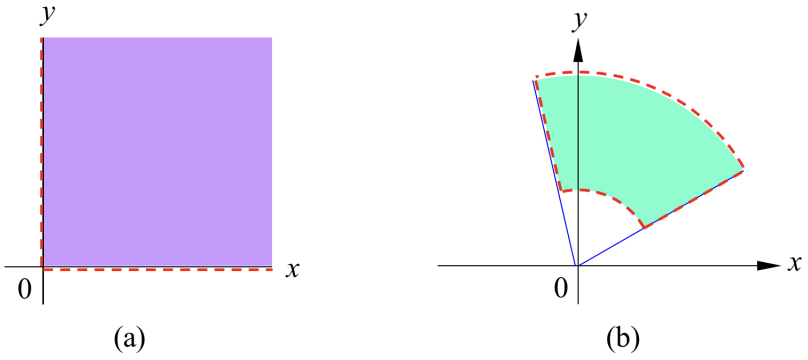


Figure 6.40: The region $\mathfrak{D} = \{(r\cos\theta, r\sin\theta) \,|\, r_1 < r < r_2, \theta_1 < \theta < \theta_2\}$ in the (x, y)-plane. (a) $r_1 = 0$, $r_2 = \infty$, $\theta_1 = 0$, $\theta_2 = \frac{\pi}{2}$. (b) $r_1 = 2$, $r_2 = 5$, $\theta_1 = \frac{\pi}{6}$, $\theta_2 = \frac{4\pi}{7}$.

To apply the change of variables theorem, we notice that the Jacobian is

$$\frac{\partial(x, y)}{\partial(r, \theta)} = r.$$

Thus,

$$dxdy = \frac{\partial(x, y)}{\partial(r, \theta)} drd\theta = rdrd\theta.$$

The change of variables theorem says the following.

Theorem 6.63

Let $[\alpha, \beta]$ be a closed interval such that $\beta \leq \alpha + 2\pi$. Assume that $g : [\alpha, \beta] \to \mathbb{R}$ and $h : [\alpha, \beta] \to \mathbb{R}$ are continuous functions satisfying

$$0 \leq g(\theta) \leq h(\theta) \qquad \text{for all } \alpha \leq \theta \leq \beta.$$

Let $\mathfrak{D}$ be the region in the (x, y)-plane given by

$$\mathfrak{D} = \{(r\cos\theta, r\sin\theta) \,|\, \alpha \leq \theta \leq \beta, g(\theta) \leq r \leq h(\theta)\}.$$

If $f : \mathfrak{D} \to \mathbb{R}$ is a continuous function, then

$$\int_{\mathfrak{D}} f(x, y)dxdy = \int_{\alpha}^{\beta} \int_{g(\theta)}^{h(\theta)} f(r\cos\theta, r\sin\theta)rdrd\theta. \tag{6.11}$$

Proof

Let

$$\mathcal{U} = \{(r, \theta) \,|\, \alpha \leq \theta \leq \beta, g(\theta) \leq r \leq h(\theta)\}.$$

Then $\mathcal{U}$ is a compact Jordan measurable set in $\mathbb{R}^2$.
If $\beta < \alpha + 2\pi$, take any α_0 such that

$$\alpha_0 < \alpha < \beta < \alpha_0 + 2\pi.$$

For example, we can take

$$\alpha_0 = \alpha - \frac{2\pi - (\beta - \alpha)}{2}.$$

If we also have

$$g(\theta) > 0 \qquad \text{for all } \theta \in [a, b],$$

then $\mathcal{U}$ is contained in the set

$$\mathcal{O}_{\alpha_0} = \{(r, \theta) \,|\, r > 0, \alpha_0 < \theta < \alpha_0 + 2\pi\},$$

and $\mathbf{\Phi}(\mathcal{U}) = \mathfrak{D}$. Applying the change of variables theorem to the mapping $\mathbf{\Phi} : \mathcal{O}_{\alpha_0} \to \mathbb{R}^2$ gives the desired formula (6.11) immediately.

If $g(\theta) = 0$ for some $\theta \in [\alpha, \beta]$, then we consider the set

$$\mathcal{U}_\varepsilon = \{(r, \theta) \,|\, \alpha \le \theta \le \beta, g(\theta) + \varepsilon \le r \le h(\theta) + \varepsilon\}, \quad \text{where } \varepsilon > 0.$$

It is contained in $\mathcal{O}_{\alpha_0}$. Using boundedness of the continuous functions $g : [\alpha, \beta] \to \mathbb{R}$, $h : [\alpha, \beta] \to \mathbb{R}$ and $f : \mathfrak{D} \to \mathbb{R}$, it is easy to show that

$$\int_{\mathfrak{D}} f(x, y)dxdy = \lim_{\varepsilon \to 0^+} \int_{\mathbf{\Phi}(\mathcal{U}_\epsilon)} f(x, y)dxdy.$$

By the change of variables formula, we have

$$\int_{\mathbf{\Phi}(\mathcal{U}_\epsilon)} f(x, y)dxdy = \int_\alpha^\beta \int_{g(\theta)+\varepsilon}^{h(\theta)+\varepsilon} f(r\cos\theta, r\sin\theta)rdrd\theta.$$

Taking the $\varepsilon \to 0^+$ limit yields again the desired formula (6.11).
The last case we have to consider is when $\beta = \alpha + 2\pi$. The technicality is that $\mathcal{U}$ is not contained in any of the $\mathcal{O}_{\alpha_0}$ restricted to which the mapping $\mathbf{\Phi}(r, \theta) = (r\cos\theta, r\sin\theta)$ is a smooth change of variables. Instead of taking limits, there is an alternative way to resolve the problem. We write $\mathcal{U} = \mathcal{U}_1 \cup \mathcal{U}_2$, where

$$\mathcal{U}_1 = \{(r, \theta) \,|\, \alpha \le \theta \le \alpha + \pi, g(\theta) \le r \le h(\theta)\},$$
$$\mathcal{U}_2 = \{(r, \theta) \,|\, \alpha + \pi \le \theta \le \alpha + 2\pi, g(\theta) \le r \le h(\theta)\},$$

We have shown that the change of variables formula (6.11) is valid for $\mathcal{U}_1$ and $\mathcal{U}_2$. Apply the additivity theorem, we find that

$$\begin{aligned}\int_{\mathfrak{D}} f(x, y)dxdy &= \int_{\mathbf{\Psi}(\mathcal{U}_1)} f(x, y)dxdy + \int_{\mathbf{\Psi}(\mathcal{U}_2)} f(x, y)dxdy \\ &= \int_\alpha^{\alpha+\pi} \int_{g(\theta)}^{h(\theta)} f(r\cos\theta, r\sin\theta)rdrd\theta \\ &\quad + \int_{\alpha+\pi}^{\alpha+2\pi} \int_{g(\theta)}^{h(\theta)} f(r\cos\theta, r\sin\theta)rdrd\theta \\ &= \int_\alpha^\beta \int_{g(\theta)}^{h(\theta)} f(r\cos\theta, r\sin\theta)rdrd\theta.\end{aligned}$$

Namely, the formula (6.11) is still valid when $\beta = \alpha + 2\pi$.

Let us give a geometric explanation for the Jacobian

$$\frac{\partial(x,y)}{\partial(r,\theta)} = r, \quad \text{where } x = r\cos\theta, y = r\sin\theta.$$

Assume that

$$\theta_1 < \theta_2 < \theta_1 + 2\pi \quad \text{and} \quad 0 < r_1 < r_2.$$

Let

$$\mathcal{D} = \{(r\cos\theta, r\sin\theta) \,|\, r_1 \le r \le r_2, \theta_1 \le \theta \le \theta_2\}.$$

The area bounded between the circles $x^2+y^2 = r_1^2$ and $x^2+y^2 = r_2^2$ is $\pi(r_2^2 - r_1^2)$. By rotational symmetry of the circle, the area of $\mathcal{D}$ is

$$\Delta A = \pi(r_2^2 - r_1^2) \times \frac{\theta_2 - \theta_1}{2\pi} = \overline{r}\Delta r\Delta\theta,$$

where

$$\overline{r} = \frac{r_1 + r_2}{2}, \quad \Delta r = r_2 - r_1 \quad \text{and} \quad \Delta\theta = \theta_2 - \theta_1.$$

When $\Delta r \to 0$, then

$$\Delta A \sim r\Delta r\Delta\theta,$$

where $r \sim r_1 \sim r_2$.

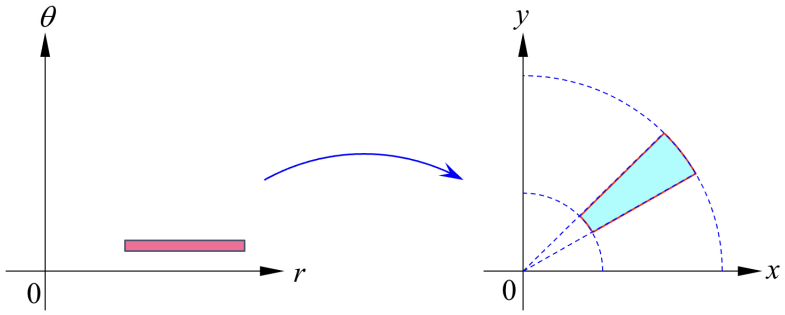


Figure 6.41: The rectangle $[r_1, r_2] \times [\theta_1, \theta_2]$ in the (r, θ)-plane is mapped to the region $\{(r\cos\theta, r\sin\theta) \,|\, r_1 \le r \le r_2, \theta_1 \le \theta \le \theta_2\}$ in the (x, y)-plane.

Now let us look at some examples.

Example 6.54

Evaluate the integral $\int_{\mathfrak{D}} (2x^2 + y^2)dxdy$, where

$$\mathfrak{D} = \left\{(x, y) \,|\, x \geq 0, y \geq 0, 4x^2 + 9y^2 \leq 36\right\}.$$

Solution

Making a change of variables $x = 3u$, $y = 2v$, the Jacobian is

$$\frac{\partial(x, y)}{\partial(u, v)} = \det \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix} = 6.$$

Then

$$\begin{aligned} \int_{\mathfrak{D}} (2x^2 + y^2)dxdy &= \int_{\mathcal{U}} \left(18u^2 + 4v^2\right) \left|\frac{\partial(x, y)}{\partial(u, v)}\right| dudv \\ &= 6 \int_{\mathcal{U}} \left(18u^2 + 4v^2\right) dudv, \end{aligned}$$

where $\mathcal{U}$ is the region

$$\mathcal{U} = \left\{(u, v) \,|\, u \geq 0, v \geq 0, u^2 + v^2 \leq 1\right\}.$$

Since $\mathcal{U}$ is symmetric if we interchange u and v, we find that

$$\int_{\mathcal{U}} u^2 dudv = \int_{\mathcal{U}} v^2 dudv = \frac{1}{2} \int_{\mathcal{U}} (u^2 + v^2)dudv.$$

Using polar coordinates $u = r\cos\theta$, $v = r\sin\theta$, we find that

$$\int_{\mathcal{U}} (u^2 + v^2)dudv = \int_0^{\frac{\pi}{2}} \int_0^1 r^2 \times rdrd\theta = \frac{\pi}{2} \int_0^1 r^3 dr = \frac{\pi}{8}.$$

Therefore,

$$\int_{\mathfrak{D}} (2x^2 + y^2)dxdy = \frac{6 \times (18 + 4)}{2} \int_{\mathcal{U}} (u^2 + v^2)dudv = \frac{33\pi}{4}.$$

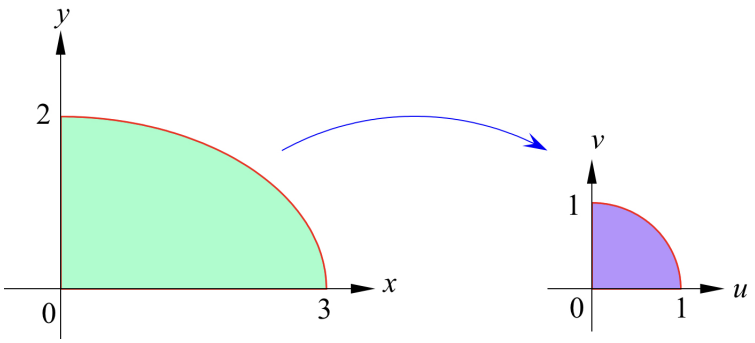


Figure 6.42: The regions $\mathfrak{D} = \{(x, y) \,|\, x \geq 0, y \geq 0, 4x^2 + 9y^2 \leq 36\}$ and $\mathcal{U} = \{(u, v) \,|\, u \geq 0, v \geq 0, u^2 + v^2 \leq 1\}$.

Example 6.55

Find the volume of the solid bounded between the surface $z = x^2 + y^2$ and the plane $z = 9$.

Solution

The solid $\mathcal{S}$ bounded between the surface $z = x^2 + y^2$ and the plane $z = 9$ can be expressed as

$$\mathcal{S} = \left\{(x, y, z) \,|\, (x, y) \in \mathfrak{D}, x^2 + y^2 \leq z \leq 9\right\},$$

where

$$\mathfrak{D} = \left\{(x, y) \,|\, x^2 + y^2 \leq 9\right\}.$$

Since $\mathfrak{D}$ is a closed ball, it is a Jordan measurable set. The volume of $\mathcal{S}$ is the integral of the constant function $\chi_{\mathcal{S}} : \mathcal{S} \to \mathbb{R}$. It is a continuous function. By Fubini's theorem,

$$\text{vol}\,(\mathcal{S}) = \int_{\mathfrak{D}} \left(\int_{x^2+y^2}^{9} dz \right) dxdy = \int_{\mathfrak{D}} (9 - x^2 - y^2) dxdy$$

Using polar coordinatets, we have

$$\text{vol}\,(\mathcal{S}) = \int_0^{2\pi} \int_0^3 (9 - r^2) r dr d\theta = 2\pi \left[\frac{9r^2}{2} - \frac{r^4}{4} \right]_0^3 = \frac{81\pi}{2}.$$

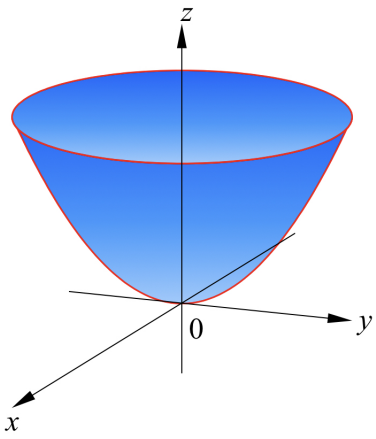


Figure 6.43: The solid bounded between the surface $z = x^2 + y^2$ and the plane $z = 9$.

Example 6.56

Let a be a positive number. Find the volume of ball

$$B = \left\{(x, y, z) \,|\, x^2 + y^2 + z^2 \leq a^2\right\}.$$

Solution

Let

$$\mathfrak{D} = \left\{(x, y) \,|\, x^2 + y^2 \leq a^2\right\}.$$

Then the ball B can be decsribed as

$$B = \left\{(x, y, z) \,|\, (x, y) \in \mathfrak{D}, -\sqrt{a^2 - x^2 - y^2} \leq z \leq \sqrt{a^2 - x^2 - y^2}\right\}.$$

Thus, by Fubini's theorem, its volume is

$$\begin{aligned} \text{vol}\,(B) &= \int_B dxdydz = \int_{\mathfrak{D}} \left(\int_{-\sqrt{a^2-x^2-y^2}}^{\sqrt{a^2-x^2-y^2}} dz \right) dxdy \\ &= \int_{\mathfrak{D}} 2\sqrt{a^2 - x^2 - y^2} dxdy \end{aligned}$$

Using polar coordinates,

$$\text{vol}\,(B) = 2\int_0^{2\pi}\int_0^a \sqrt{a^2-r^2}\, r dr d\theta = 4\pi \int_0^a r\sqrt{a^2-r^2}\, dr.$$

Let $u = a^2 - r^2$. Then $du = -2rdr$. When $r = 0$, $u = a^2$. When $r = a$, $u = 0$. Therefore,

$$\text{vol}\,(B) = 2\pi \int_0^{a^2} u^{\frac{1}{2}} du = 2\pi \left[\frac{2}{3}u^{\frac{3}{2}}\right]_0^{a^2} = \frac{4\pi}{3}a^3.$$

Example 6.57

Let a be a positive number, and let α be a number in the interval $(0, \frac{\pi}{2})$. Find the volume of the solid E bounded between the sphere

$$S = \left\{(x,y,z)\,|\,x^2+y^2+z^2 = a^2\right\}$$

and the cone

$$C = \left\{(x,y,z)\,|\,z = \cot\alpha\sqrt{x^2+y^2}\right\}.$$

Solution

The surfaces S and C intersect at the points (x,y,z) satisfying

$$(x^2+y^2)(1+\cot^2\alpha) = a^2.$$

Namely,

$$x^2+y^2 = a^2\sin^2\alpha.$$

Therefore,

$$E = \left\{(x,y,z)\,|\,(x,y)\in\mathfrak{D}, \cot\alpha\sqrt{x^2+y^2} \le z \le \sqrt{a^2-x^2-y^2}\right\},$$

where

$$\mathfrak{D} = \left\{(x,y)\,|\,x^2+y^2 \le a^2\sin^2\alpha\right\}.$$

Using Fubini's theorem and polar coordinates, we find that

$$\begin{aligned}\text{vol}\,(E) &= \int_E dxdydz = \int_{\mathfrak{D}} \left(\int_{\cot\alpha\sqrt{x^2+y^2}}^{\sqrt{a^2-x^2-y^2}} dz \right) dxdy \\ &= \int_0^{2\pi} \int_0^{a\sin\alpha} \left(\sqrt{a^2-r^2} - r\cot\alpha \right) r dr d\theta\end{aligned}$$

Using a change of variables $u = a^2 - r^2$, we find that

$$\begin{aligned}\int_0^{a\sin\alpha} r\sqrt{a^2-r^2}dr &= \frac{1}{2}\int_{a^2\cos^2\alpha}^{a^2} u^{\frac{1}{2}}du \\ &= \frac{1}{3}\left[u^{\frac{3}{2}}\right]_{a^2\cos^2\alpha}^{a^2} = \frac{a^3}{3}\left(1-\cos^3\alpha\right).\end{aligned}$$

On the other hand

$$\begin{aligned}\int_0^{a\sin\alpha} r^2\cot\alpha\, dr &= \cot\alpha\left[\frac{r^3}{3}\right]_0^{a\sin\alpha} \\ &= \frac{a^3}{3}\frac{\cos\alpha}{\sin\alpha}\sin^3\alpha = \frac{a^3}{3}(\cos\alpha - \cos^3\alpha).\end{aligned}$$

Therefore, the volume of E is

$$\text{vol}\,(E) = \frac{2\pi a^3}{3}(1-\cos\alpha).$$

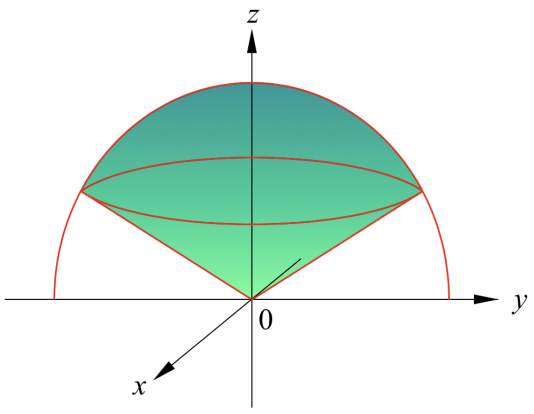


Figure 6.44: The solid bounded between a sphere and a cone.

6.5.3 Spherical Coordinates

Now we consider the spherical coordinates, which is an alternative coordinate system for $\mathbb{R}^3$. Consider the mapping $\mathbf{\Psi} : \mathbb{R}^3 \to \mathbb{R}^3$ given by

$$\mathbf{\Psi}(\rho, \phi, \theta) = (x, y, z) = (\rho \sin\phi \cos\theta, \rho \sin\phi \sin\theta, \rho \cos\phi).$$

Namely,

$$\begin{aligned} x &= \rho \sin\phi \cos\theta, \\ y &= \rho \sin\phi \sin\theta, \\ z &= \rho \cos\phi. \end{aligned}$$

Let V be the set

$$V = \{(\rho, \phi, \theta) \,|\, \rho > 0, 0 \le \phi \le \pi, 0 \le \theta < 2\pi\}.$$

Given $\mathbf{u} = (x, y, z) \in \mathbb{R}^3 \setminus \{(0, 0, 0)\}$, we claim that there is a unique $(\rho, \phi, \theta) \in V$ such that $\mathbf{\Psi}(\rho, \phi, \theta) = (x, y, z)$. This triple (ρ, ϕ, θ) is called a spherical coordinates of the point $\mathbf{u} = (x, y, z)$. It is easy to see that

$$\rho = \sqrt{x^2 + y^2 + z^2} = \|\mathbf{u}\|$$

is the distance from the point $\mathbf{u} = (x, y, z)$ to the origin. If we let

$$\phi = \cos^{-1} \frac{z}{\rho},$$

ϕ satisfies $0 \le \phi \le \pi$, and

$$\langle \mathbf{u}, \mathbf{e}_3 \rangle = z = \rho \cos\phi = \|\mathbf{u}\| \cos\phi.$$

Thus, geometrically, ϕ is the angle the vector from $\mathbf{0}$ to $\mathbf{u}$ makes with the positive z-axis. Let W be the (x, y)-plane in $\mathbb{R}^3$. Then

$$(x, y, 0) = \text{proj}_W \mathbf{u}.$$

Let

$$r = \rho \sin\phi.$$

Then

$$r = \sqrt{\rho^2 - \rho^2 \cos^2\phi} = \sqrt{\rho^2 - z^2} = \sqrt{x^2 + y^2}.$$

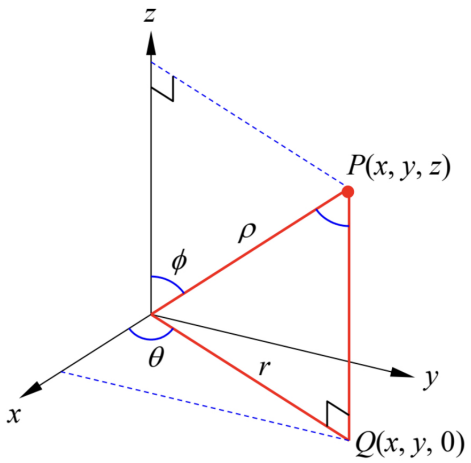


Figure 6.45: Spherical Coordinates.

Thus, $\theta \in [0, 2\pi)$ is uniquely determined so that

$$x = r\cos\theta, \quad y = r\sin\theta.$$

Equivalently, (r, θ) is the polar coordinates of the point (x, y) in $\mathbb{R}^2 \setminus \{(0, 0)\}$. Hence, we find that the map $\mathbf{\Psi} : V \to \mathbb{R}^3$,

$$\mathbf{\Psi}(\rho, \phi, \theta) = (x, y, z) = (\rho\sin\phi\cos\theta, \rho\sin\phi\sin\theta, \rho\cos\phi)$$

is one-to-one on the set V, and the range is $\mathbb{R}^3 \setminus \{(0, 0, 0)\}$. However, the inverse map is not continuous.

Let us calculate the derivative matrix of $\mathbf{\Psi} : \mathbb{R}^3 \to \mathbb{R}^3$. We find that

$$\mathbf{D\Psi}(\rho, \phi, \theta) = \begin{bmatrix} \sin\phi\cos\theta & \rho\cos\phi\cos\theta & -\rho\sin\phi\sin\theta \\ \sin\phi\sin\theta & \rho\cos\phi\sin\theta & \rho\sin\phi\cos\theta \\ \cos\phi & -\rho\sin\phi & 0 \end{bmatrix}.$$

Therefore, the Jacobian $\det \mathbf{D\Psi}(\rho, \phi, \theta)$ is

$$\begin{aligned} \frac{\partial(x, y, z)}{\partial(\rho, \phi, \theta)} &= \cos\phi \times \det \begin{bmatrix} \rho\cos\phi\cos\theta & -\rho\sin\phi\sin\theta \\ \rho\cos\phi\sin\theta & \rho\sin\phi\cos\theta \end{bmatrix} \\ &\quad + \rho\sin\phi \times \det \begin{bmatrix} \sin\phi\cos\theta & -\rho\sin\phi\sin\theta \\ \sin\phi\sin\theta & \rho\sin\phi\cos\theta \end{bmatrix} \\ &= \rho^2\cos^2\phi\sin\phi + \rho^2\sin^3\phi = \rho^2\sin\phi. \end{aligned}$$

This shows that $\mathbf{D}\boldsymbol{\Psi}(\rho,\phi,\theta)$ is invertible if and only if $\rho \neq 0$ and $\sin\phi \neq 0$, if and only if (x,y,z) does not lie on the z axis. Thus, for any real number α, if $\mathcal{O}_\alpha$ is the open set

$$\mathcal{O}_\alpha = \{(\rho,\phi,\theta) \,|\, \rho > 0, 0 < \phi < \pi, \alpha < \theta < \alpha + 2\pi\},$$

then $\boldsymbol{\Psi} : \mathcal{O}_\alpha \to \mathbb{R}^3$ is a smooth change of variables.

The change of variables theorem gives the following.

Theorem 6.64

Let $[\alpha,\beta]$ and $[\delta,\eta]$ be a closed intervals such that

$$\beta \leq \alpha + 2\pi \quad \text{and} \quad 0 \leq \delta < \eta \leq \pi,$$

and let $\mathbf{I} = [\delta,\eta] \times [\alpha,\beta]$. Assume that the functions $g : \mathbf{I} \to \mathbb{R}$ and $h : \mathbf{I} \to \mathbb{R}$ satisfy $0 \leq g(\phi,\theta) \leq h(\phi,\theta)$ for all $(\phi,\theta) \in \mathbf{I}$, let $\mathfrak{D}$ be the region in $\mathbb{R}^3$ defined by

$$\mathfrak{D} = \{(\rho\sin\phi\cos\theta, \rho\sin\phi\sin\theta, \rho\cos\phi) \,|\, (\phi,\theta) \in \mathbf{I}, g(\phi,\theta) \leq \rho \leq h(\phi,\theta)\}.$$

If $f : \mathfrak{D} \to \mathbb{R}$ is a continuous function, then

$$\begin{aligned}&\int_{\mathfrak{D}} f(x,y,z)dxdydz \\ &= \int_\alpha^\beta \int_\delta^\eta \int_{g(\phi,\theta)}^{h(\phi,\theta)} f(\rho\sin\phi\cos\theta, \rho\sin\phi\sin\theta, \rho\cos\phi)\rho^2\sin\phi\, d\rho d\phi d\theta.\end{aligned}$$

Again, if $\beta < \alpha + 2\pi$, $\delta > 0$, $\eta < \pi$ and $g(\phi,\theta) > 0$ for all $(\phi,\theta) \in \mathbf{I}$, this is just a direct consequence of the general change of variables theorem. The rest can be argued by taking limits.

The results of Example 6.57 can be used to give a hindsight about the Jacobian

$$\frac{\partial(x,y,z)}{\partial(\rho,\phi,\theta)} = \rho^2\sin\phi$$

that appears in the change from spherical coordinates to rectangular coordinates. Consider the rectangle $\mathbf{I} = [\rho_1,\rho_2] \times [\phi_1,\phi_2] \times [\theta_1,\theta_2]$ in the (ρ,ϕ,θ) space, where $\theta_2 < \theta_1 + 2\pi$, and for simplicity, assume that $0 < \phi_1 < \phi_2 < \frac{\pi}{2}$. Under the

mapping

$$\mathbf{\Psi}(\rho, \phi, \theta) = (\rho \sin\phi \cos\theta, \rho \sin\phi \sin\theta, \rho \cos\phi),$$

$\mathbf{\Psi}(\mathbf{I})$ is a wedge in the solid E in $\mathbb{R}^3$ bounded between the spheres $x^2 + y^2 + z^2 = \rho_1^2$, $x^2 + y^2 + z^2 = \rho_2^2$, and the cones $z = \cot\phi_1\sqrt{x^2 + y^2}$, $z = \cot\phi_2\sqrt{x^2 + y^2}$. Since E has a rotational symmetry with respect to θ,

$$\Delta V = \text{vol}\,(\mathbf{\Psi}(\mathbf{I})) = \frac{\theta_2 - \theta_1}{2\pi}\,\text{vol}\,(E).$$

Using inclusion and exclusion principle, the result of Example 6.57 gives

$$\begin{aligned}\text{vol}\,(E) &= \frac{2\pi}{3}\rho_2^3(1 - \cos\phi_2) - \frac{2\pi}{3}\rho_1^3(1 - \cos\phi_2)\\ &\quad - \frac{2\pi}{3}\rho_2^3(1 - \cos\phi_1) + \frac{2\pi}{3}\rho_1^3(1 - \cos\phi_1)\\ &= \frac{2\pi}{3}(\rho_2^3 - \rho_1^3)(\cos\phi_1 - \cos\phi_2).\end{aligned}$$

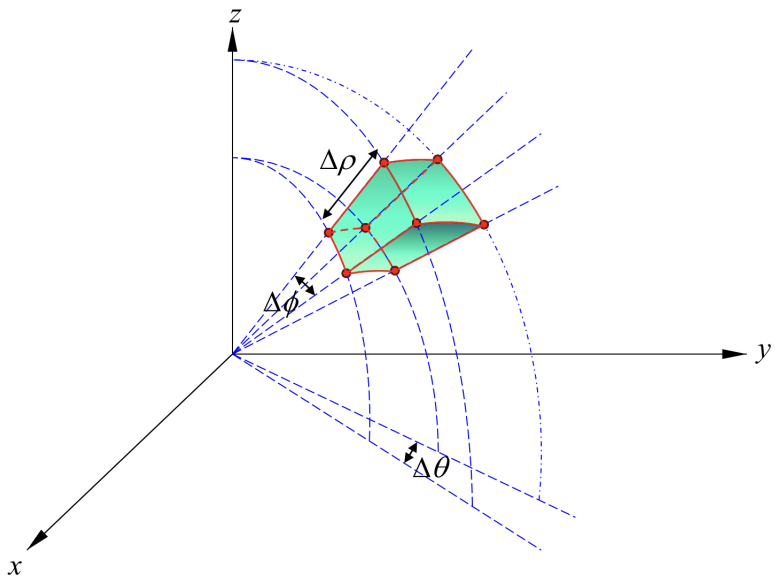


Figure 6.46: Volume change under spherical coordinates.

By mean value theorem there is a $\overline{\rho} \in (\rho_1, \rho_2)$ and a $\overline{\phi} \in (\phi_1, \phi_2)$ such that

$$\rho_2^3 - \rho_1^3 = 3\overline{\rho}^2\Delta\rho \quad \text{and} \quad \cos\phi_1 - \cos\phi_2 = \sin\overline{\phi}\Delta\phi,$$

where

$$\Delta\rho = \rho_2 - \rho_1, \qquad \Delta\phi = \phi_2 - \phi_1.$$

Let $\Delta\theta = \theta_2 - \theta_1$. Then we find that

$$\Delta V = \overline{\rho}^2 \sin\overline{\phi}\,\Delta\rho\Delta\phi\Delta\theta.$$

This gives an interpretation of the Jacobian $\rho^2 \sin\phi$.

Let us look at an example of applying spherical coordinates.

Example 6.58

Compute the integral $\int_E (x^2 + 4z)dxdydz$, where

$$E = \left\{(x, y, z) \,|\, x^2 + y^2 + z^2 \leq 9, z \geq 0\right\}.$$

Solution

Let B be the sphere

$$B = \left\{(x, y, z) \,|\, x^2 + y^2 + z^2 \leq 9\right\}.$$

By symmetry,

$$\int_E x^2 dxdydz = \frac{1}{2}\int_B x^2 dxdydz = \frac{1}{6}\int_B (x^2 + y^2 + z^2)dxdydz.$$

Using spherical coordinates, we have

$$\begin{aligned}\int_E x^2 dxdydz &= \frac{1}{6}\int_0^{2\pi}\int_0^{\pi}\int_0^3 \rho^2 \times \rho^2 \sin\phi\, d\rho d\phi d\theta \\ &= \frac{\pi}{3}\left[-\cos\phi\right]_0^{\pi}\left[\frac{\rho^5}{5}\right]_0^3 = \frac{162\pi}{5}.\end{aligned}$$

On the other hand,

$$\begin{aligned}\int_E z dxdydz &= \int_0^{2\pi}\int_0^{\frac{\pi}{2}}\int_0^3 \rho\cos\phi \times \rho^2 \sin\phi\, d\rho d\phi d\theta \\ &= 2\pi\left[-\frac{\cos^2\phi}{2}\right]_0^{\frac{\pi}{2}}\left[\frac{\rho^4}{4}\right]_0^3 = \frac{81\pi}{4}.\end{aligned}$$

Therefore,

$$\int_E (x^2 + 4z)dxdydz = \frac{162\pi}{5} + 81\pi = \frac{567\pi}{5}.$$

In the example above, we have used the symmetry of the region E to avoid some complicated computations. Another example is the following.

Example 6.59

Let a be a positive number. Evaluate the integral $\int_E x^4 dxdydz$, where

$$E = \left\{(x, y, z) \,|\, x \geq 0, y \geq 0, z \geq 0, x^2 + y^2 + z^2 \leq a^2\right\}.$$

Solution

The expression of the z variable in terms of the spherical coordinates is considerably simpler than the x and y variables. By symmetry, we have

$$\int_E x^4 dxdydz = \int_E z^4 dxdydz.$$

Thus, using spherical coordinates, we find that

$$\begin{aligned}\int_E x^4 dxdydz &= \int_0^{\frac{\pi}{2}} \int_0^{\frac{\pi}{2}} \int_0^a \rho^4 \cos^4\phi \times \rho^2 \sin\phi \, d\rho d\phi d\theta \\ &= \frac{\pi}{2}\left[-\frac{\cos^5\theta}{5}\right]_0^{\frac{\pi}{2}} \left[\frac{\rho^7}{7}\right]_0^a = \frac{\pi a^7}{70}.\end{aligned}$$

6.5.4 Other Examples

Example 6.60

Let $\mathfrak{D}$ be the region

$$\mathfrak{D} = \left\{(x, y) \,|\, y > 0, 4 \leq x^2 - y^2 \leq 9, 3 \leq xy \leq 7\right\}.$$

Compute the integral

$$\int_{\mathfrak{D}} (x^3 y + xy^3)dxdy.$$

Solution

Let $\mathcal{O} = \{(x, y) \,|\, y > 0\}$, and let $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^2$ be the mapping

$$\mathbf{\Psi}(x, y) = (x^2 - y^2, xy).$$

If (x_1, y_1) and (x_2, y_2) are points in $\mathcal{O}$ such that $\mathbf{\Psi}(x_1, y_1) = \mathbf{\Psi}(x_2, y_2)$, then

$$x_1^2 - y_1^2 = x_2^2 - y_2^2 \quad \text{and} \quad x_1 y_1 = x_2 y_2.$$

Let $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$. Then

$$z_1^2 = (x_1 + iy_1)^2 = (x_2 + iy_2)^2 = z_2^2.$$

This implies that $z_2 = \pm z_1$. Thus, $y_2 = \pm y_1$. Since y_1 and y_2 are positive, we find that $y_1 = y_2$. Since $x_1 y_1 = x_2 y_2$, we then deduce that $x_1 = x_2$. Hence, $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^2$ is one-to-one. Since it is a polynomial mapping, it is continuously differentiable. Since

$$\mathbf{D\Psi}(x, y) = \begin{bmatrix} 2x & -2y \\ y & x \end{bmatrix}, \qquad \det \mathbf{D\Psi}(x, y) = 2(x^2 + y^2),$$

we find that $\det \mathbf{D\Psi}(x, y) \neq 0$ for all $(x, y) \in \mathcal{O}$. This implies that $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^2$ is a smooth change of variables. Let $u = x^2 - y^2$, $v = xy$. The Jacobian is

$$\frac{\partial(u, v)}{\partial(x, y)} = 2(x^2 + y^2).$$

Notice that

$$\mathbf{\Psi}(\mathfrak{D}) = \{(u, v) \,|\, 4 \leq u \leq 9, 3 \leq v \leq 7\}.$$

Therefore,

$$\begin{aligned} \int_{\mathfrak{D}} xy(x^2 + y^2) dx dy &= \frac{1}{2} \int_{\mathfrak{D}} xy \frac{\partial(u, v)}{\partial(x, y)} dx dy \\ &= \frac{1}{2} \int_{\mathbf{\Psi}(\mathfrak{D})} v du dv = \frac{1}{2} \int_3^7 \int_4^9 v du dv = 50. \end{aligned}$$

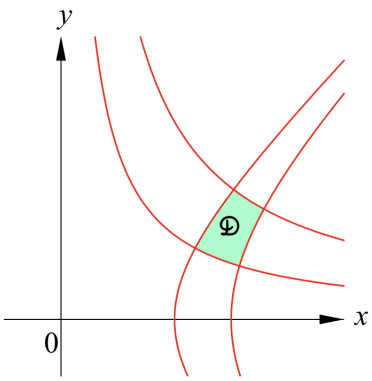


Figure 6.47: The region $\mathfrak{D} = \{(x, y) \,|\, y > 0, 4 \le x^2 - y^2 \le 9, 3 \le xy \le 7\}$.

Remark 6.11 Hyperspherical Coordinates

For any $n \ge 4$, the hyperspherical coorfinates in $\mathbb{R}^n$ are the coordinates $(r, \theta_1, \ldots, \theta_{n-1})$ such that

$$\begin{aligned} x_1 &= r\cos\theta_1, \\ x_2 &= r\sin\theta_1\cos\theta_2, \\ x_3 &= r\sin\theta_1\sin\theta_2\cos\theta_3, \\ &\vdots \\ x_{n-1} &= r\sin\theta_1\cdots\sin\theta_{n-2}\cos\theta_{n-1} \\ x_n &= r\sin\theta_1\cdots\sin\theta_{n-2}\sin\theta_{n-1}. \end{aligned} \tag{6.12}$$

Here

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

If

$$V = (0, \infty) \times [0, \pi)^{n-2} \times [0, 2\pi),$$

there is a one-to-one correspondence between $(r, \theta_1, \ldots, \theta_{n-1})$ in V and $(x_1, \ldots, x_n)$ in $\mathbb{R}^n \setminus \{\mathbf{0}\}$ given by (6.12). One can show that the Jacobian of this transformation is

$$\frac{\partial(x_1, x_2, \ldots, x_n)}{\partial(r, \theta_1, \ldots, \theta_{n-1})} = r^{n-1}\sin^{n-2}\theta_1\cdots\sin\theta_{n-2}.$$

Exercises 6.5

Question 1

Let

$$\mathfrak{D} = \left\{(x, y) \,|\, 4(x+1)^2 + 9(y-2)^2 \leq 144\right\}.$$

Evaluate the integral $\displaystyle\int_{\mathfrak{D}} (2x + 3y)dxdy$.

Question 2

Evaluate the integral

$$\int_{\mathfrak{D}} \frac{x+y}{(x-2y+8)^2} dxdy,$$

where

$$\mathfrak{D} = \{(x, y) \,|\, |x| + 2|y| \leq 7\}.$$

Question 3

Evaluate the integral $\displaystyle\int_{\mathfrak{D}} (x^2 - xy + y^2)dxdy$, where

$$\mathfrak{D} = \left\{(x, y) \,|\, x \geq 0, x^2 + 9y^2 \leq 36\right\}.$$

Question 4

Find the volume of the solid bounded between the surface $z = x^2 + y^2$ and the surface $x^2 + y^2 + z^2 = 20$.

Question 5

Let a, b and c be positive numbers, and let E be the solid

$$E = \left\{(x, y, z) \;\middle|\; \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} \leq 1\right\}.$$

Evaluate $\displaystyle\int_E x^2 dxdydz$.

Question 6

Let a, b and c be positive numbers, and let E be the solid

$$E = \left\{ (x, y, z) \;\middle|\; x \geq 0, \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} \leq 1 \right\}.$$

Evaluate $\displaystyle\int_E x^6 dxdydz$.

Question 7

Let $\mathfrak{D}$ be the region

$$\mathfrak{D} = \left\{ (x, y) \,|\, x > 0, 1 \leq x^2 - y^2 \leq 25, 1 \leq xy \leq 6 \right\}.$$

Compute the integral

$$\int_{\mathfrak{D}} \frac{x^4 - y^4}{xy} dxdy.$$

Question 8

Let $\mathfrak{D}$ be the region

$$\mathfrak{D} = \left\{ (x, y) \,|\, 5x^2 - 2xy + 10y^2 \leq 9 \right\},$$

and let $\mathbf{\Psi} : \mathbb{R}^2 \to \mathbb{R}^2$ be the mapping defined by

$$\mathbf{\Psi}(x, y) = (2x + y, x - 3y).$$

(a) Explain why $\mathbf{\Psi} : \mathbb{R}^2 \to \mathbb{R}^2$ is a smooth change of variables.

(b) Find $\mathbf{\Psi}(\mathfrak{D})$.

(c) Compute the integral $\displaystyle\int_{\mathfrak{D}} \frac{8}{5x^2 - 2xy + 10y^2 + 16} \, dxdy$.

6.6 Proof of the Change of Variables Theorem

In this section, we give a complete proof of the change of variables theorem, which we restate here.

Theorem 6.65 The Change of Variables Theorem

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^n$ be a smooth change of variables. If $\mathfrak{D}$ is a Jordan measurable set such that its closure $\overline{\mathfrak{D}}$ is contained in $\mathcal{O}$, then $\mathbf{\Psi}(\mathfrak{D})$ is Jordan measurable, and for any function $f : \mathbf{\Psi}(\mathfrak{D}) \to \mathbb{R}$ that is bounded and continuous, we have

$$\int_{\mathbf{\Psi}(\mathfrak{D})} f(\mathbf{x}) d\mathbf{x} = \int_{\mathfrak{D}} f\left(\mathbf{\Psi}(\mathbf{x})\right) \left|\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})\right| d\mathbf{x}.$$

Among the assertions in the theorem, we will first establish the following.

Theorem 6.66

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^n$ be a smooth change of variables. If $\mathfrak{D}$ is a Jordan measurable set such that its closure $\overline{\mathfrak{D}}$ is contained in $\mathcal{O}$, then $\mathbf{\Psi}(\mathfrak{D})$ is also Jordan measurable.

A special case of the change of variables theorem is when $f : \mathbf{\Psi}(\mathfrak{D}) \to \mathbb{R}$ is the characteristic function of $\mathbf{\Psi}(\mathfrak{D})$. This gives the change of volume theorem.

Theorem 6.67 The Change of Volume Theorem

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^n$ be a smooth change of variables. If $\mathfrak{D}$ is a Jordan measurable set such that its closure $\overline{\mathfrak{D}}$ is contained in $\mathcal{O}$, then

$$\text{vol}\left(\mathbf{\Psi}(\mathfrak{D})\right) = \int_{\mathfrak{D}} \left|\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})\right| d\mathbf{x}.$$

In the following, let us give some remarks about the statements in the theorem, and outline the plan of the proof.

The Change of Variables Theorem

1. The first step is to prove Theorem 6.66 which asserts that $\boldsymbol{\Psi}(\mathfrak{D})$ is Jordan measurable. To do this, we first show that a smooth change of variables sets up a one-to-one correspondence between the open sets in the domain and the range. This basically follows from inverse function theorem.

2. Since $f : \boldsymbol{\Psi}(\mathfrak{D}) \to \mathbb{R}$ is continuous and bounded, if $\boldsymbol{\Psi}(\mathfrak{D})$ is Jordan measurable, $f : \boldsymbol{\Psi}(\mathfrak{D}) \to \mathbb{R}$ is Riemann integrable.

3. Let $g : \mathcal{O} \to \mathbb{R}$ be the function

$$g(\mathbf{x}) = |\det \mathbf{D}\boldsymbol{\Psi}(\mathbf{x})|.$$

   Since $\boldsymbol{\Psi} : \mathcal{O} \to \mathbb{R}^n$ is continuously differentiable, the function $\mathbf{D}\boldsymbol{\Psi} : \mathcal{O} \to \mathbb{R}^{n^2}$ is continuous. Since determinant and absolute value are continuous functions, $g : \mathcal{O} \to \mathbb{R}$ is a continuous function. Since $\overline{\mathfrak{D}}$ is a compact set contained in $\mathcal{O}$, $g : \overline{\mathfrak{D}} \to \mathbb{R}$ is bounded.

4. Since $\boldsymbol{\Psi} : \mathcal{O} \to \mathbb{R}^n$ is continuous, and the functions $f : \boldsymbol{\Psi}(\mathfrak{D}) \to \mathbb{R}$ and $g : \mathfrak{D} \to \mathbb{R}$ are continuous and bounded, the function $h : \mathfrak{D} \to \mathbb{R}$,

$$h(\mathbf{x}) = f\left(\boldsymbol{\Psi}(\mathbf{x})\right) g(\mathbf{x}) = f\left(\boldsymbol{\Psi}(\mathbf{x})\right) |\det \mathbf{D}\boldsymbol{\Psi}(\mathbf{x})|$$

   is continuous and bounded. Hence, it is Riemann integrable.

5. To prove the change of variables theorem, we will first prove the change of volume theorem. This is the most technical part of the proof.

6. To prove the change of volume theorem, we first consider the case where $\boldsymbol{\Psi} : \mathbb{R}^n \to \mathbb{R}^n$ is an invertible linear transformation. In this case, the theorem says that if $\mathfrak{D}$ is a Jordan measurable set, and $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{T}(\mathbf{x}) = A\mathbf{x}$ is an invertible linear transformation, then

$$\text{vol}\left(\mathbf{T}(\mathfrak{D})\right) = |\det A| \,\text{vol}\left(\mathfrak{D}\right). \tag{6.13}$$

7. To prove (6.13), we first consider the case where $\mathfrak{D} = \mathbf{I}$ is a closed rectangle. This is an easy consequence of the fact that the volume of a parallelepiped spanned by the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is equal to $|\det A|$, where A is the matrix with $\mathbf{v}_1, \ldots, \mathbf{v}_n$ as column vectors. This was proved in Appendix B.

8. After proving the change of volume theorem, we will prove the change of variables theorem for the special case where $\mathfrak{D} = \mathbf{I}$ is a closed rectangle first. The general theorem then follows by some simple analysis argument.

We begin by the following proposition which says that a smooth change of variables maps open sets to open sets.

Proposition 6.68

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^n$ be a smooth change of variables. Then for any open set $\mathcal{D}$ that is contained in $\mathcal{O}$, $\mathbf{\Psi}(\mathcal{D})$ is open in $\mathbb{R}^n$. In particular, $\mathbf{\Psi}(\mathcal{O})$ is an open subset of $\mathbb{R}^n$.

Proof

Given that $\mathcal{D}$ is an open subset of $\mathbb{R}^n$, let $\mathcal{W} = \mathbf{\Psi}(\mathcal{D})$. We want to show that $\mathcal{W}$ is an open set. If $\mathbf{y}_0$ is a point in $\mathcal{W}$, there is an $\mathbf{x}_0$ in $\mathcal{D}$ such that $\mathbf{y}_0 = \mathbf{\Psi}(\mathbf{x}_0)$. Since $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^n$ is continuously differentiable and $\mathbf{D}\mathbf{\Psi}(\mathbf{x}_0)$ is invertible, we can apply inverse function theorem to conclude that there is an open set $\mathcal{U}_0$ containing $\mathbf{x}_0$ such that $\mathbf{\Psi}(\mathcal{U}_0)$ is also open, and $\mathbf{\Psi}^{-1} : \mathbf{\Psi}(\mathcal{U}_0) \to \mathcal{U}_0$ is continuously differentiable. Let $\mathcal{U} = \mathcal{U}_0 \cap \mathcal{D}$. Then $\mathcal{U}$ is an open subset of $\mathcal{D}$ and $\mathcal{U}_0$. It follows that $\mathcal{V} = \mathbf{\Psi}(\mathcal{U}) = (\mathbf{\Psi}^{-1})^{-1}(\mathcal{U})$ is an open subset of $\mathbb{R}^n$ that is contained in $\mathcal{W} = \mathbf{\Psi}(\mathcal{D})$. Notice that $\mathcal{V}$ is an open set that contains $\mathbf{y}_0$. Thus, we have shown that every point in $\mathcal{W}$ has a neighbourhood that lies in $\mathcal{W}$. This proves that $\mathcal{W}$ is an open set.

The following proposition says that the inverse of a smooth change of variables is also a smooth change of variables.

Proposition 6.69

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^n$ be a smooth change of variables. Then $\mathbf{\Psi}^{-1} : \mathbf{\Psi}(\mathcal{O}) \to \mathbb{R}^n$ is also a smooth change of variables.

Proof

By Proposition 6.68, $\mathbf{\Psi}(\mathcal{O})$ is an open set. By default, $\mathbf{\Psi}^{-1} : \mathbf{\Psi}(\mathcal{O}) \to \mathbb{R}^n$ is one-to-one. As in the proof of Proposition 6.68, the inverse function theorem implies that it is continuously differentiable. If $\mathbf{x}_0 = \mathbf{\Psi}^{-1}(\mathbf{y}_0)$, inverse function theorem says that

$$\mathbf{D}\mathbf{\Psi}^{-1}(\mathbf{y}_0) = \mathbf{D}\mathbf{\Psi}(\mathbf{x}_0)^{-1}.$$

The inverse of an invertible matrix is invertible. Hence, for any $\mathbf{y}_0$ in $\mathbf{\Psi}(\mathcal{O})$, $\mathbf{D}\mathbf{\Psi}^{-1}(\mathbf{y}_0)$ is invertible. These prove that $\mathbf{\Psi}^{-1} : \mathbf{\Psi}(\mathcal{O}) \to \mathbb{R}^n$ is a smooth change of variables.

Remark 6.12 Homeomorphisms and Diffeomorphisms

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^n$ be a continuous injective map such that $\mathbf{\Psi}(\mathcal{O})$ is open, and the inverse map $\mathbf{\Psi}^{-1} : \mathbf{\Psi}(\mathcal{O}) \to \mathcal{O}$ is continuous. Then we say that $\mathbf{\Psi} : \mathcal{O} \to \mathbf{\Psi}(\mathcal{O})$ is a *homeomorphism*. A homeomorphism sets up a one-to-one correspondence between open sets in $\mathcal{O}$ and open sets in $\mathbf{\Psi}(\mathcal{O})$.
If $\mathbf{\Psi} : \mathcal{O} \to \mathbf{\Psi}(\mathcal{O})$ is a homeomorphism and both the maps $\mathbf{\Psi} : \mathcal{O} \to \mathbf{\Psi}(\mathcal{O})$ and $\mathbf{\Psi}^{-1} : \mathbf{\Psi}(\mathcal{O}) \to \mathcal{O}$ are continuously differentiable, then we say that $\mathbf{\Psi} : \mathcal{O} \to \mathbf{\Psi}(\mathcal{O})$ is a *diffeomorphism*. Proposition 6.68 and Proposition 6.69 imply that a continuous change of variables is a diffeomorphism.
A map of the form $\mathbf{\Psi} : \mathbb{R}^n \to \mathbb{R}^n$,

$$\mathbf{\Psi}(\mathbf{x}) = \mathbf{y}_0 + A(\mathbf{x} - \mathbf{x}_0),$$

where $\mathbf{x}_0$ and $\mathbf{y}_0$ are points in $\mathbb{R}^n$ and A is an invertible matrix, is a diffeomorphism.

Now we can prove the following which is essential for the proof of Theorem

6.66.

Theorem 6.70

Assume that $\mathcal{O}$ and $\mathcal{U}$ are open subsets of $\mathbb{R}^n$, and $\mathbf{\Psi} : \mathcal{O} \to \mathcal{U}$ is a homeomorphism. If $\mathfrak{D}$ is a subset of $\mathcal{O}$ such that $\overline{\mathfrak{D}}$ is also contained in $\mathcal{O}$, then

$$\operatorname{int} \mathbf{\Psi}(\mathfrak{D}) = \mathbf{\Psi}(\operatorname{int} \mathfrak{D}), \qquad \overline{\mathbf{\Psi}(\mathfrak{D})} = \mathbf{\Psi}(\overline{\mathfrak{D}}).$$

Thus,

$$\partial \mathbf{\Psi}(\mathfrak{D}) = \mathbf{\Psi}(\partial \mathfrak{D}).$$

Proof

The interior of a set A is an open set that contains all the open set that is contained in A. By Remark 6.12, there is a one-to-one correspondence between the open sets that are contained in $\mathfrak{D}$ and the open sets that are contained in $\mathbf{\Psi}(\mathfrak{D})$. Therefore,

$$\operatorname{int} \mathbf{\Psi}(\mathfrak{D}) = \mathbf{\Psi}(\operatorname{int} \mathfrak{D}).$$

Since $\overline{\mathfrak{D}}$ is a compact set and $\mathbf{\Psi} : \mathcal{O} \to \mathcal{U}$ is continuous, $\mathbf{\Psi}(\overline{\mathfrak{D}})$ is a compact set. Therefore, $\mathbf{\Psi}(\overline{\mathfrak{D}})$ is a closed set that contains $\mathbf{\Psi}(\mathfrak{D})$. This implies that

$$\overline{\mathbf{\Psi}(\mathfrak{D})} \subset \mathbf{\Psi}(\overline{\mathfrak{D}}). \tag{6.14}$$

Since $\mathbf{\Psi}^{-1} : \mathcal{U} \to \mathcal{O}$ is also continuous, the same argument gives

$$\overline{\mathfrak{D}} = \overline{\mathbf{\Psi}^{-1}\left(\mathbf{\Psi}(\mathfrak{D})\right)} \subset \mathbf{\Psi}^{-1}(\overline{\mathbf{\Psi}(\mathfrak{D})}).$$

This implies that

$$\mathbf{\Psi}\left(\overline{\mathfrak{D}}\right) \subset \overline{\mathbf{\Psi}(\mathfrak{D})}. \tag{6.15}$$

Eq. (6.14) and (6.15) give

$$\overline{\mathbf{\Psi}(\mathfrak{D})} = \mathbf{\Psi}(\overline{\mathfrak{D}}).$$

The last assertion follows from the fact that for any set A, $\overline{A}$ is a disjoint union of $\operatorname{int} A$ and ∂A.

Recall that a set $\mathfrak{D}$ in $\mathbb{R}^n$ has Jordan content zero if and only if for every $\varepsilon > 0$,

$\mathfrak{D}$ can be covered by finitely many cubes $Q_1, \ldots, Q_k$, such that

$$\sum_{j=1}^{k} \text{vol}\,(Q_j) < \varepsilon.$$

The next proposition gives a control of the size of the cube under a smooth change of variables.

Proposition 6.71

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^n$ be a smooth change of variables. If $Q_{\mathbf{c},r}$ is a cube with center at $\mathbf{c}$ and side length $2r$, then $\mathbf{\Psi}(Q_{\mathbf{c},r})$ is contained in the cube $Q_{\mathbf{\Psi}(\mathbf{c}),\lambda r}$, where

$$\lambda = \max_{1\le i\le n} \max_{\mathbf{x}\in Q_{\mathbf{c},r}} \sum_{j=1}^{n} \left| \frac{\partial \Psi_i}{\partial x_j}(\mathbf{x}) \right|.$$

Therefore,

$$\text{vol}\,\left(\mathbf{\Psi}(Q_{\mathbf{c},r})\right) \le \lambda^n \text{vol}\,(Q_{\mathbf{c},r}).$$

Remark 6.13

Note that since $Q_{\mathbf{c},r}$ is a compact set and $\dfrac{\partial \Psi_i}{\partial x_j}(\mathbf{x})$ is continuous for all $1 \le i, j \le n$,

$$\max_{\mathbf{x}\in Q_{\mathbf{c},r}} \sum_{j=1}^{n} \left| \frac{\partial \Psi_i}{\partial x_j}(\mathbf{x}) \right|$$

exists.

Proof of Proposition 6.71

Notice that $\mathbf{u} \in Q_{\mathbf{c},r}$ if and only if

$$|u_i - c_i| \le r \qquad \text{for each } 1 \le i \le n.$$

Let $\mathbf{d} = \mathbf{\Psi}(\mathbf{c})$. Given $\mathbf{v} = \mathbf{\Psi}(\mathbf{u})$ with $\mathbf{u} \in Q_{\mathbf{c},r}$, we want to show that $\mathbf{v}$ is in $Q_{\mathbf{d},\lambda r}$, or equivalently,

$$|v_i - d_i| \le \lambda r \qquad \text{for each } 1 \le i \le n.$$

This is basically an application of mean value theorem. The set $Q_{\mathbf{c},r}$ is convex and the map $\Psi_i : \mathcal{O} \to \mathbb{R}$ is continuously differentiable. Mean value theorem says that there is a point $\mathbf{x}$ in $Q_{\mathbf{c},r}$ such that

$$v_i - d_i = \Psi_i(\mathbf{u}) - \Psi_i(\mathbf{c}) = \sum_{j=1}^{n} \frac{\partial \Psi_i}{\partial x_j}(\mathbf{x})(u_j - c_j).$$

Therefore,

$$\begin{aligned} |v_i - d_i| &\leq \sum_{j=1}^{n} \left| \frac{\partial \Psi_i}{\partial x_j}(\mathbf{x}) \right| |u_j - c_j| \leq r \sum_{j=1}^{n} \left| \frac{\partial \Psi_i}{\partial x_j}(\mathbf{x}) \right| \\ &\leq r \max_{\mathbf{x} \in Q_{\mathbf{c},r}} \sum_{j=1}^{n} \left| \frac{\partial \Psi_i}{\partial x_j}(\mathbf{u}) \right| \leq \lambda r. \end{aligned}$$

This proves that $\mathbf{\Psi}(Q_{\mathbf{c},r})$ is contained in $Q_{\mathbf{\Psi}(\mathbf{c}),\lambda r}$. The last assertion in the proposition about the volumes is obvious.

Now we prove Theorem 6.66.

Proof of Theorem 6.66

Since $\overline{\mathfrak{D}}$ is a compact set that is contain in the open set $\mathcal{O}$, Theorem 3.36 says that there is a positive number d and a compact set C such that $\mathfrak{D} \subset C \subset \mathcal{O}$, and any point in $\mathbb{R}^n$ that has a distance less than d from a point in $\overline{\mathfrak{D}}$ lies in C.

Since $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^n$ is continuously differentiable, for all $1 \leq i, j \leq n$, $\dfrac{\partial \Psi_i}{\partial x_j} : C \to \mathbb{R}$ is a continuous function. Since C is a compact set, for each $1 \leq i \leq n$, the function

$$\sum_{j=1}^{n} \left| \frac{\partial \Psi_i}{\partial x_j} \right| (\mathbf{x})$$

has a maximum on C. Hence,

$$\lambda = \max_{1 \leq i \leq n} \max_{\mathbf{x} \in C} \sum_{j=1}^{n} \left| \frac{\partial \Psi_i}{\partial x_j} \right| (\mathbf{x})$$

exists.

Since $\mathfrak{D}$ is Jordan measurable, $\partial\mathfrak{D}$ has Jordan content zero. Since C contains $\overline{\mathfrak{D}}$, it contains $\partial\mathfrak{D}$. Given $\varepsilon > 0$, there exist cubes $Q_1, Q_2, \ldots, Q_k$, each of which intersects $\mathfrak{D}$, and such that

$$\partial\mathfrak{D} \subset \bigcup_{j=1}^{n} Q_j \quad \text{and} \quad \sum_{j=1}^{k} \text{vol}\,(Q_j) < \frac{\varepsilon}{\lambda^n}.$$

Since a uniformly regular partition of a cube will divide the cube into cubes, we can also assume that each of the cubes Q_j, $1 \leq j \leq k$ has diameter less than d. This implies that each Q_j, $1 \leq j \leq k$ is contained in C. For $1 \leq j \leq k$, let l_j be the side length of Q_j. Proposition 6.71 says that $\mathbf{\Psi}(Q_j)$ is contained in a cube $\widetilde{Q}_j$ with side length λl_j. Therefore,

$$\partial\mathbf{\Psi}(\mathfrak{D}) = \mathbf{\Psi}(\partial\mathfrak{D}) \subset \bigcup_{j=1}^{k} \widetilde{Q}_j,$$

$$\text{and} \qquad \sum_{j=1}^{k} \text{vol}\,(\widetilde{Q}_j) \leq \lambda^n \sum_{j=1}^{k} \text{vol}\,(Q_j) < \varepsilon.$$

This shows that $\partial\mathbf{\Psi}(\mathfrak{D})$ has Jordan content zero. Hence, $\mathbf{\Psi}(\mathfrak{D})$ is a Jordan measurable set.

To prove the change of volume formula, the crucial thing is to first prove the special case where $\mathfrak{D} = \mathbf{I}$ is a rectangle, and $\mathbf{\Psi}$ is an invertible linear transformation. In Appendix B, we prove the following theorem which gives the volume of a parallelepiped.

Theorem 6.72

Let $\mathscr{P}$ be a parallelepiped in $\mathbb{R}^n$ spanned by the linearly independent vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$. Then the volume of $\mathscr{P}$ is equal to $|\det A|$, where A is the matrix whose column vectors are $\mathbf{v}_1, \ldots, \mathbf{v}_n$.

We then use this to deduce the following special case of the change of volume formula.

Theorem 6.73

Let $\mathbf{I}$ be a closed rectangle in $\mathbb{R}^n$, and let $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{T}(\mathbf{x}) = A\mathbf{x}$ be an invertible linear transformation. Then

$$\text{vol}\,(\mathbf{T}(\mathbf{I})) = |\det A|\,\text{vol}\,(\mathbf{I}).$$

Using this, we can prove the more general change of volume formula for a Jordan measurable set under an invertible linear transformation.

Theorem 6.74

If $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{T}(\mathbf{x}) = A\mathbf{x}$ is an invertible linear transformation, and $\mathfrak{D}$ is a Jordan measurable set, then $\mathbf{T}(\mathfrak{D})$ is also Jordan measurable and

$$\text{vol}\,(\mathbf{T}(\mathfrak{D})) = |\det A|\,\text{vol}\,(\mathfrak{D}).$$

Proof

Since $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$ is invertible, $\det A \neq 0$. The fact that the set $\mathbf{T}(\mathfrak{D})$ is Jordan measurable follows from Theorem 6.66. Let $\mathbf{I}$ be a closed rectangle that contains $\mathfrak{D}$. Since $\mathfrak{D}$ is Jordan measurable, the characteristic function $\chi_{\mathfrak{D}} : \mathbf{I} \to \mathbb{R}$ is Riemann integrable.
Given $\varepsilon > 0$, there is a partition $\mathbf{P}$ of $\mathbf{I}$ such that

$$U(\chi_{\mathfrak{D}}, \mathbf{P}) - \int_{\mathbf{I}} \chi_{\mathfrak{D}} < \frac{\varepsilon}{|\det A|}.$$

Hence,

$$U(\chi_{\mathfrak{D}}, \mathbf{P}) < \text{vol}\,(\mathfrak{D}) + \frac{\varepsilon}{|\det A|}.$$

Let

$$\mathscr{A} = \{\mathbf{J} \in \mathcal{J}_{\mathbf{P}} \,|\, \mathbf{J} \cap \mathfrak{D} \neq \emptyset\}.$$

Then

$$\sum_{\mathbf{J} \in \mathscr{A}} \text{vol}\,(\mathbf{J}) = U(\chi_{\mathfrak{D}}, \mathbf{P}) < \text{vol}\,(\mathfrak{D}) + \frac{\varepsilon}{|\det A|}.$$

Notice that

$$\mathfrak{D} \subset \bigcup_{\mathbf{J} \in \mathscr{A}} \mathbf{J}.$$

Therefore,

$$\mathbf{T}(\mathfrak{D}) \subset \bigcup_{\mathbf{J}\in\mathscr{A}} \mathbf{T}(\mathbf{J}).$$

For each rectangle $\mathbf{J}$, $\mathbf{T}(\mathbf{J})$ is a parallelepiped. For any two distinct rectangles in $\mathscr{A}$, they are disjoint or intersect at a set that has Jordan content zero. Therefore, additivity theorem implies that the set K defined as

$$K = \bigcup_{\mathbf{J}\in\mathscr{A}} \mathbf{T}(\mathbf{J})$$

is Jordan measurable, and

$$\text{vol}\,(K) = \sum_{\mathbf{J}\in\mathscr{A}} \text{vol}\,(\mathbf{T}(\mathbf{J})) = |\det A| \sum_{\mathbf{J}\in\mathscr{A}} \text{vol}\,(\,\mathbf{J}) < |\det A|\,\text{vol}\,(\mathfrak{D}) + \varepsilon.$$

Since $\mathbf{T}(\mathfrak{D}) \subset K$, we find that

$$\text{vol}\,(\mathbf{T}(\mathfrak{D})) \leq \text{vol}\,(K) < |\det A|\,\text{vol}\,(\mathfrak{D}) + \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, we conclude that

$$\text{vol}\,(\mathbf{T}(\mathfrak{D})) \leq |\det A|\,\text{vol}\,(\mathfrak{D}). \tag{6.16}$$

Since $\mathbf{T}^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ is also an invertible linear transformation, we find that

$$\begin{aligned} \text{vol}\,(\mathfrak{D}) &= \text{vol}\,\left(\mathbf{T}^{-1}\left(\mathbf{T}(\mathfrak{D})\right)\right) \\ &\leq |\det A^{-1}|\,\text{vol}\,(\mathbf{T}(\mathfrak{D})) = \frac{1}{|\det A|}\text{vol}\,(\mathbf{T}(\mathfrak{D})). \end{aligned} \tag{6.17}$$

Eq. (6.16) and (6.17) together give

$$\text{vol}\,(\mathbf{T}(\mathfrak{D})) = |\det A|\,\text{vol}\,(\mathfrak{D}).$$

Recall that by identitying an $n \times n$ matrix $A = [a_{ij}]$ as a point in $\mathbb{R}^{n^2}$, we have defined the norm of A as

$$\|A\| = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2}.$$

Besides the triangle inequality, this norm also satisfies the following identity.

Lemma 6.75

If $A = [a_{ij}]$ and $B = [b_{ij}]$ are $n \times n$ matrices, then

$$\|AB\| \leq \|A\| \, \|B\|.$$

Proof

Let $[c_{ij}] = C = AB$. Then for any $1 \leq i, j \leq n$,

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}.$$

By Cauchy-Schwarz inequality,

$$c_{ij}^2 \leq \left(\sum_{k=1}^{n} a_{ik}^2 \right) \left(\sum_{l=1}^{n} b_{lj}^2 \right).$$

Therefore,

$$\|C\|^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}^2 \leq \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\sum_{k=1}^{n} a_{ik}^2 \right) \left(\sum_{l=1}^{n} b_{lj}^2 \right) = \|A\|^2 \|B\|^2.$$

This proves that

$$\|AB\| = \|C\| \leq \|A\| \, \|B\|.$$

Now we prove the change of volume formula, which is the most technical part.

Proof of Theorem 6.67

Given the smooth change of variables $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^n$, let $g : \mathcal{O} \to \mathbb{R}$ be the continuous function

$$g(\mathbf{x}) = |\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})|.$$

We want to show that if $\mathfrak{D}$ is a Jordan measurable set such that its closure $\overline{\mathfrak{D}}$ is contained in $\mathcal{O}$, then

$$\text{vol } (\mathbf{\Psi}(\mathfrak{D})) = \int_{\mathfrak{D}} |\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})| d\mathbf{x} = \int_{\mathfrak{D}} g(\mathbf{x}) d\mathbf{x} = \mathcal{I}.$$

We will first prove that

$$\text{vol}\left(\boldsymbol{\Psi}(\mathfrak{D})\right) \leq \mathcal{I}.$$

By Theorem 6.66, $\boldsymbol{\Psi}(\mathfrak{D})$ is Jordan measurable. By Theorem 6.49, its closure $\overline{\boldsymbol{\Psi}(\mathfrak{D})} = \boldsymbol{\Psi}(\overline{\mathfrak{D}})$ is also Jordan measurable, and

$$\text{vol}\left(\boldsymbol{\Psi}(\mathfrak{D})\right) = \text{vol}\left(\overline{\boldsymbol{\Psi}(\mathfrak{D})}\right) = \text{vol}\left(\boldsymbol{\Psi}(\overline{\mathfrak{D}})\right).$$

On the other hand, since $\overline{\mathfrak{D}} \setminus \mathfrak{D}$ has Jordan content zero,

$$\int_{\mathfrak{D}} g(\mathbf{x})d\mathbf{x} = \int_{\overline{\mathfrak{D}}} g(\mathbf{x})d\mathbf{x}.$$

Hence, we can assume from the beginning that $\mathfrak{D} = \overline{\mathfrak{D}}$, or equivalently, $\mathfrak{D}$ is closed.

As in the proof of Theorem 6.66, Theorem 3.36 says that there is a positive number d and a compact set C such that $\mathfrak{D} \subset C \subset \mathcal{O}$, and any point in $\mathbb{R}^n$ that has a distance less than d from a point in $\mathfrak{D}$ lies in C. On the compact set C, the function $g : C \to \mathbb{R}$ is continuous. By extreme value theorem, there are points $\mathbf{u}$ and $\mathbf{v}$ in C such that $g(\mathbf{u}) \leq g(\mathbf{x}) \leq g(\mathbf{v})$ for all $\mathbf{x} \in C$. Let $m_g = g(\mathbf{u})$ and $M_g = g(\mathbf{v})$. Then $m_g > 0$ and

$$m_g \leq g(\mathbf{x}) \leq M_g \qquad \text{for all } \mathbf{x} \in C.$$

On the other hand, the function $\mathbf{D}\boldsymbol{\Psi}^{-1} : C \to \mathbb{R}^{n^2}$ is continuous on the compact set C. Hence, it is bounded. Namely, there is a positive number M_h such that

$$\|\mathbf{D}\boldsymbol{\Psi}^{-1}(\mathbf{x})\| \leq M_h \qquad \text{for all } \mathbf{x} \in C.$$

Let L be a positive number such that C is contained in the cube $\mathbf{I} = [-L, L]^n$. Let $\check{g} : \mathbf{I} \to \mathbb{R}$ be the zero extension of $g : \mathfrak{D} \to \mathbb{R}$. For each positive integer k, let $\mathbf{P}_k$ be the uniformly regular partition of $\mathbf{I}$ into k^n rectangles. Then $\lim_{k\to\infty} |\mathbf{P}_k| = 0$. Therefore,

$$\lim_{k\to\infty} U(\check{g}, \mathbf{P}_k) = \int_{\mathbf{I}} \check{g}(\mathbf{x})d\mathbf{x} = \int_{\mathfrak{D}} g(\mathbf{x})d\mathbf{x} = \mathcal{I},$$

and

$$\lim_{k\to\infty} \left(U(\chi_{\mathfrak{D}}, \mathbf{P}_k) - L(\chi_{\mathfrak{D}}, \mathbf{P}_k)\right) = 0.$$

The compactness of C implies that the continuous functions $\mathbf{D}\boldsymbol{\Psi} : C \to \mathbb{R}^{n^2}$ and $g : C \to \mathbb{R}$ are uniformly continuous. Given $\varepsilon > 0$, there exists a $\delta_1 > 0$ such that if $\mathbf{u}$ and $\mathbf{v}$ are points in C with $\|\mathbf{u} - \mathbf{v}\| < \delta_1$, then

$$\|\mathbf{D}\boldsymbol{\Psi}(\mathbf{u}) - \mathbf{D}\boldsymbol{\Psi}(\mathbf{v})\| < \frac{\varepsilon}{M_h n} \quad \text{and} \quad |g(\mathbf{u}) - g(\mathbf{v})| < m_g \varepsilon.$$

Let $\delta = \min\{d, \delta_1\}$. There is a positive integer K such that for all $k \geq K$, $|\mathbf{P}_k| < \delta$,

$$U(\check{g}, \mathbf{P}_k) < \mathcal{I} + \varepsilon \quad \text{and} \quad U(\chi_{\mathfrak{D}}, \mathbf{P}_k) - L(\chi_{\mathfrak{D}}, \mathbf{P}_k) < \frac{\varepsilon}{M_g}.$$

Consider the partition $\mathbf{P} = \mathbf{P}_K$. Since $\mathbf{I}$ is a cube and $\mathbf{P}$ is a uniformly regular partition of $\mathbf{I}$, each rectangle in the partition $\mathbf{P}$ is also a cube, all with the same side length $2r$. Denote by $\mathscr{A}$ and $\mathscr{B}$ the sets

$$\mathscr{A} = \{\mathbf{J} \in \mathcal{J}_{\mathbf{P}} \,|\, \mathbf{J} \cap \mathfrak{D} \neq \emptyset\}, \qquad \mathscr{B} = \{\mathbf{J} \in \mathcal{J}_{\mathbf{P}} \,|\, \mathbf{J} \subset \mathfrak{D}\}.$$

$\mathscr{A}$ is a finite collection of cubes, and it contains the collection $\mathscr{B}$. By definition,

$$L(\chi_{\mathfrak{D}}, \mathbf{P}) = \sum_{\mathbf{J}\in\mathscr{B}} \text{vol}\,(\mathbf{J}), \qquad U(\chi_{\mathfrak{D}}, \mathbf{P}) = \sum_{\mathbf{J}\in\mathscr{A}} \text{vol}\,(\mathbf{J}).$$

Therefore,

$$\sum_{\mathbf{J}\in\mathscr{A}\setminus\mathscr{B}} \text{vol}\,(\mathbf{J}) < \frac{\varepsilon}{M_g}.$$

After renaming, we can assume that

$$\mathscr{A} = \{Q_\beta \,|\, 1 \leq \beta \leq s\},$$

where $Q_\beta = Q_{\mathbf{c}_\beta, r}$ is a cube with center at $\mathbf{c}_\beta$ and side length $2r$. By the definition of $\mathscr{A}$,

$$\mathfrak{D} \subset \bigcup_{\beta=1}^{s} Q_\beta.$$

Therefore,

$$\mathbf{\Psi}(\mathfrak{D}) \subset \bigcup_{\beta=1}^{s} \mathbf{\Psi}\left(Q_\beta\right).$$

Fixed a $1 \le \beta \le s$. Since Q_β intersects $\mathfrak{D}$ and

$$\operatorname{diam} Q_\beta = |\mathbf{P}| < \delta \le d,$$

we find that Q_β is contained in C. Define the invertible linear transformation $\mathbf{T}_\beta : \mathbb{R}^n \to \mathbb{R}^n$ by $\mathbf{T}_\beta(\mathbf{x}) = A_\beta \mathbf{x}$, where $A_\beta = \mathbf{D\Psi}(\mathbf{c}_\beta)$. Then let $\mathbf{\Phi}_\beta : Q_\beta \to \mathbb{R}^n$ be the map $\mathbf{\Phi}_\beta = \mathbf{T}_\beta^{-1} \circ \mathbf{\Psi} : Q_\beta \to \mathbb{R}^n$. By chain rule,

$$\mathbf{D\Phi}_\beta(\mathbf{x}) = \mathbf{D}\mathbf{T}_\beta^{-1}(\mathbf{\Psi}(\mathbf{x}))\, \mathbf{D\Psi}(\mathbf{x}) = A_\beta^{-1}\, \mathbf{D\Psi}(\mathbf{x}) \quad \text{for all } \mathbf{x} \in Q_\beta.$$

Therefore, for $\mathbf{x} \in Q_\beta$,

$$\mathbf{D\Phi}_\beta(\mathbf{x}) - I_n = \mathbf{D\Psi}^{-1}(\mathbf{\Psi}(\mathbf{c}_\beta))\left(\mathbf{D\Psi}(\mathbf{x}) - \mathbf{D\Psi}(\mathbf{c}_\beta)\right).$$

Since $\|\mathbf{x} - \mathbf{c}_\beta\| \le \operatorname{diam} Q_\beta < \delta \le \delta_1$, Lemma 6.75 implies that

$$\|\mathbf{D\Phi}_\beta(\mathbf{x}) - I_n\| \le \left\|\mathbf{D\Psi}^{-1}(\mathbf{\Psi}(\mathbf{c}_\beta))\right\| \, \|\mathbf{D\Psi}(\mathbf{x}) - \mathbf{D\Psi}(\mathbf{c}_\beta)\| < \frac{\varepsilon}{n}.$$

This implies that if $i \ne j$,

$$\left|\frac{\partial(\Phi_\beta)_i}{\partial x_j}(\mathbf{x})\right| < \frac{\varepsilon}{n} \qquad \text{for all } \mathbf{x} \in Q_\beta;$$

while

$$\left|\frac{\partial(\Phi_\beta)_i}{\partial x_i}(\mathbf{x})\right| < 1 + \frac{\varepsilon}{n} \qquad \text{for all } \mathbf{x} \in Q_\beta.$$

Hence,

$$\lambda_\beta = \max_{1 \le i \le n} \max_{\mathbf{x} \in Q_\beta} \sum_{j=1}^{n} \left|\frac{\partial(\Phi_\beta)_i}{\partial x_j}(\mathbf{x})\right| \le 1 + \varepsilon.$$

Since $\mathbf{\Psi} = \mathbf{T}_\beta \circ \mathbf{\Phi}_\beta$, by Theorem 6.74 and Proposition 6.71, we have

$$\operatorname{vol}\left(\mathbf{\Psi}(Q_\beta) = |\det A_\beta| \operatorname{vol}\left(\mathbf{\Phi}_\beta(Q_\beta)\right) \le |\det A_\beta|\, \lambda_\beta^n \operatorname{vol}\left(Q_\beta\right).\right.$$

Summing over β, we find that

$$\operatorname{vol}\left(\mathbf{\Psi}(\mathfrak{D})\right) \leq \sum_{\beta=1}^{s} \operatorname{vol}\left(\mathbf{\Psi}\left(Q_\beta\right)\right) \leq (1+\varepsilon)^n \sum_{\beta=1}^{s} |\det \mathbf{D}\mathbf{\Psi}(\mathbf{c}_\beta)| \operatorname{vol}\left(Q_\beta\right).$$

We divide the sum into a sum over those Q_β in $\mathscr{B}$ and a sum over those Q_β in $\mathscr{A} \setminus \mathscr{B}$. For the sum over those in $\mathscr{B}$, we find that

$$\sum_{Q_\beta \in \mathscr{B}} |\det \mathbf{D}\mathbf{\Psi}(\mathbf{c}_\beta)| \operatorname{vol}\left(Q_\beta\right) \leq U(\check{g}, \mathbf{P}) < \mathcal{I} + \varepsilon.$$

For the sum over those Q_β in $\mathscr{A} \setminus \mathscr{B}$,

$$\sum_{Q_\beta \in \mathscr{A}\setminus\mathscr{B}} |\det \mathbf{D}\mathbf{\Psi}(\mathbf{c}_\beta)| \operatorname{vol}\left(Q_\beta\right) \leq M_g \sum_{Q_\beta \in \mathscr{A}\setminus\mathscr{B}} \operatorname{vol}\left(Q_\beta\right) < \varepsilon.$$

Hence,

$$\operatorname{vol}\left(\mathbf{\Psi}(\mathfrak{D})\right) \leq (1+\varepsilon)^n \left(\mathcal{I} + 2\varepsilon\right).$$

Since $\varepsilon > 0$ is arbitrary, taking the limit $\varepsilon \to 0^+$, we find that

$$\operatorname{vol}\ \left(\mathbf{\Psi}(\mathfrak{D})\right) \leq \mathcal{I} = \int_{\mathfrak{D}} |\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})|\, d\mathbf{x}.$$

This is true for any smooth change of variables $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}$ and any Jordan measurable closed subet $\mathfrak{D}$ that is contained in $\mathcal{O}$.
Now we want to prove the opposite inequality. First note that the same inequality applied to the smooth change of variables $\mathbf{\Psi}^{-1} : \mathbf{\Psi}(\mathcal{O}) \to \mathbb{R}^n$. Thus, if $\mathcal{F}$ is a closed Jordan measurable subset of $\mathbf{\Psi}(\mathcal{O})$, then

$$\operatorname{vol}\ \left(\mathbf{\Psi}^{-1}(\mathcal{F})\right) \leq \int_{\mathcal{F}} \left|\det \mathbf{D}\mathbf{\Psi}^{-1}(\mathbf{u})\right| d\mathbf{u}. \tag{6.18}$$

Using the same ε and partition $\mathbf{P}$ as above, for $1 \leq \beta \leq s$, let

$$\mathcal{F}_\beta = \mathbf{\Psi}\left(\mathfrak{D} \cap Q_\beta\right).$$

Since $\mathfrak{D}$ and Q_β are closed Jordan measurable sets, $\mathfrak{D} \cap Q_\beta$ is also a closed Jordan measurable set, and so is $\mathcal{F}_\beta$. Additivity theorem implies that

$$\mathcal{I} = \int_{\mathfrak{D}} g(\mathbf{x}) d\mathbf{x} = \sum_{\beta=1}^{s} \int_{\mathfrak{D}\cap Q_\beta} g(\mathbf{x}) d\mathbf{x}.$$

For each $1 \leq \beta \leq s$, since $\mathfrak{D} \cap Q_\beta$ is compact, there is a point $\mathbf{v}_\beta \in \mathfrak{D} \cap Q_\beta$ such that

$$g(\mathbf{x}) \leq g(\mathbf{v}_\beta) \qquad \text{for all } \mathbf{x} \in \mathfrak{D} \cap Q_\beta.$$

This gives

$$\int_{\mathfrak{D}\cap Q_\beta} g(\mathbf{x}) d\mathbf{x} \leq g(\mathbf{v}_\beta) \operatorname{vol}\left(\mathfrak{D} \cap Q_\beta\right) = g(\mathbf{v}_\beta) \operatorname{vol}\left(\boldsymbol{\Psi}^{-1}(\mathcal{F}_\beta)\right).$$

By (6.18), we find that

$$\int_{\mathfrak{D}\cap Q_\beta} g(\mathbf{x}) d\mathbf{x} \leq g(\mathbf{v}_\beta) \int_{\mathcal{F}_\beta} \left|\det \mathbf{D}\boldsymbol{\Psi}^{-1}(\mathbf{u})\right| d\mathbf{u}.$$

Again, there is a point $\mathbf{w}_\beta \in \mathfrak{D} \cap Q_\beta$ such that

$$\left|\det \mathbf{D}\boldsymbol{\Psi}^{-1}(\mathbf{u})\right| \leq \left|\det \mathbf{D}\boldsymbol{\Psi}^{-1}(\boldsymbol{\Psi}(\mathbf{w}_\beta))\right| \qquad \text{for all } \mathbf{u} \in \mathcal{F}_\beta.$$

This implies that

$$\int_{\mathcal{F}_\beta} \left|\det \mathbf{D}\boldsymbol{\Psi}^{-1}(\mathbf{u})\right| d\mathbf{u} \leq \left|\det \mathbf{D}\boldsymbol{\Psi}(\mathbf{w}_\beta))\right|^{-1} \operatorname{vol}\left(\mathcal{F}_\beta\right).$$

Hence,

$$\int_{\mathfrak{D}\cap Q_\beta} g(\mathbf{x}) d\mathbf{x} \leq \frac{g(\mathbf{v}_\beta)}{g(\mathbf{w}_\beta)} \operatorname{vol}\left(\mathcal{F}_\beta\right) = \operatorname{vol}\left(\mathcal{F}_\beta\right) \left(1 + \frac{g(\mathbf{v}_\beta) - g(\mathbf{w}_\beta)}{g(\mathbf{w}_\beta)}\right).$$

Now since $\mathbf{v}_\beta$ and $\mathbf{w}_\beta$ are in Q_β, $\|\mathbf{v}_\beta - \mathbf{w}_\beta\| < \delta_1$. Thus,

$$\left|\frac{g(\mathbf{v}_\beta) - g(\mathbf{w}_\beta)}{g(\mathbf{w}_\beta)}\right| \leq \frac{1}{m_g} \left|g(\mathbf{v}_\beta) - g(\mathbf{w}_\beta)\right| < \varepsilon.$$

This gives

$$\int_{\mathfrak{D}\cap Q_\beta} g(\mathbf{x}) d\mathbf{x} \leq (1+\varepsilon) \operatorname{vol}\left(\boldsymbol{\Psi}\left(\mathfrak{D} \cap Q_\beta\right)\right).$$

Summing over β and using additivity theorem, we find that

$$\mathcal{I} = \int_{\mathfrak{D}} g(\mathbf{x}) d\mathbf{x} \leq (1+\varepsilon) \sum_{\beta=1}^{s} \operatorname{vol}\left(\boldsymbol{\Psi}\left(\mathfrak{D} \cap Q_\beta\right)\right) = (1+\varepsilon) \operatorname{vol}\left(\boldsymbol{\Psi}(\mathfrak{D})\right).$$

Taking $\varepsilon \to 0^+$ gives the desired inequality

$$\mathcal{I} \leq \text{vol } (\mathbf{\Psi}(\mathfrak{D})) .$$

This completes the proof of the change of volume theorem.

To conclude the proof of the change of variables theorem, we need the following generalization of the mean value theorem for integrals.

Theorem 6.76 Generalized Mean Value Theorem for Integrals

Let $\mathfrak{D}$ be a compact Jordan measurable set, and let $f : \mathfrak{D} \to \mathbb{R}$ and $g : \mathfrak{D} \to \mathbb{R}$ be continuous functions. If $\mathfrak{D}$ is connected or path-connected, and $g(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathfrak{D}$, then there is a point $\mathbf{x}_0$ in $\mathfrak{D}$ such that

$$\int_{\mathfrak{D}} f(\mathbf{x})g(\mathbf{x})d\mathbf{x} = f(\mathbf{x}_0) \int_{\mathfrak{D}} g(\mathbf{x})d\mathbf{x}.$$

The proof of this generalized mean value theorem is almost the same as the mean value theorem. The latter can be considered as the special case where $g(\mathbf{x}) = 1$ for all $\mathbf{x}$ in $\mathfrak{D}$.

Proof

Since $\mathfrak{D}$ is compact and $f : \mathfrak{D} \to \mathbb{R}$ is continuous, extreme value theorem asserts that there exist points $\mathbf{u}$ and $\mathbf{v}$ in $\mathfrak{D}$ such that

$$f(\mathbf{u}) \leq f(\mathbf{x}) \leq f(\mathbf{v}) \qquad \text{for all } \mathbf{x} \in \mathfrak{D}.$$

Since $g(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathfrak{D}$, we find that

$$f(\mathbf{u})g(\mathbf{x}) \leq f(\mathbf{x})g(\mathbf{x}) \leq f(\mathbf{v})g(\mathbf{x}) \qquad \text{for all } \mathbf{x} \in \mathfrak{D}.$$

The monotonicity theorem implies that

$$f(\mathbf{u}) \int_{\mathfrak{D}} g(\mathbf{x})d\mathbf{x} \leq \int_{\mathfrak{D}} f(\mathbf{x})g(\mathbf{x})d\mathbf{x} \leq f(\mathbf{v}) \int_{\mathfrak{D}} g(\mathbf{x})d\mathbf{x}.$$

Let

$$U = \int_{\mathfrak{D}} g(\mathbf{x}) d\mathbf{x}.$$

If $U = 0$, we can take $\mathbf{x}_0$ to be any point in $\mathfrak{D}$. If $U \neq 0$, notice that

$$c = \frac{1}{U} \int_{\mathfrak{D}} f(\mathbf{x}) g(\mathbf{x}) d\mathbf{x}$$

satisfies

$$f(\mathbf{u}) \leq c \leq f(\mathbf{v}).$$

As in the proof of the mean value theorem, $\mathfrak{D}$ is connected or path-connected allows us to conclude that there is an $\mathbf{x}_0$ in $\mathfrak{D}$ such that

$$\frac{1}{U} \int_{\mathfrak{D}} f(\mathbf{x}) g(\mathbf{x}) d\mathbf{x} = c = f(\mathbf{x}_0).$$

This gives

$$\int_{\mathfrak{D}} f(\mathbf{x}) g(\mathbf{x}) d\mathbf{x} = f(\mathbf{x}_0) \int_{\mathfrak{D}} g(\mathbf{x}) d\mathbf{x}.$$

Next, we prove the special case of the change of variables theorem when $\mathfrak{D}$ is a closed rectangle.

Theorem 6.77

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and let $\mathbf{\Psi} : \mathcal{O} \to \mathbb{R}^n$ be a smooth change of variables. If $\mathbf{I}$ is a closed rectangle contained in $\mathcal{O}$, and $f : \mathbf{\Psi}(\mathbf{I}) \to \mathbb{R}$ is continuous, then

$$\int_{\mathbf{\Psi}(\mathbf{I})} f(\mathbf{x}) d\mathbf{x} = \int_{\mathbf{I}} f\left(\mathbf{\Psi}(\mathbf{x})\right) \left|\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})\right| d\mathbf{x}.$$

Proof

It is sufficient to show that for any $\varepsilon > 0$,

$$\left| \int_{\mathbf{\Psi}(\mathbf{I})} f(\mathbf{x}) d\mathbf{x} - \int_{\mathbf{I}} f\left(\mathbf{\Psi}(\mathbf{x})\right) \left|\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})\right| d\mathbf{x} \right| < \varepsilon.$$

Since $f : \mathbf{\Psi}(\mathbf{I}) \to \mathbb{R}$ and $\mathbf{\Psi} : \mathbf{I} \to \mathbb{R}^n$ are continuous, $(f \circ \mathbf{\Psi}) : \mathbf{I} \to \mathbb{R}$ is continuous. Since $\mathbf{I}$ is compact, $(f \circ \mathbf{\Psi}) : \mathbf{I} \to \mathbb{R}$ is uniformly continuous. Given $\varepsilon > 0$, there is a $\delta > 0$ such that if $\mathbf{u}$ and $\mathbf{v}$ are points in $\mathbf{I}$,

$$|f(\mathbf{\Psi}(\mathbf{u})) - f(\mathbf{\Psi}(\mathbf{v}))| < \frac{\varepsilon}{\text{vol}\,(\mathbf{\Psi}(\mathbf{I}))}.$$

Let $\mathbf{P}$ be a partition of $\mathbf{I}$ such that $|\mathbf{P}| < \delta$. By additivity theorem,

$$\int_{\mathbf{\Psi}(\mathbf{I})} f(\mathbf{x}) d\mathbf{x} = \sum_{\mathbf{J} \in \mathcal{J}_\mathbf{P}} \int_{\mathbf{\Psi}(\mathbf{J})} f(\mathbf{x}) d\mathbf{x},$$

and

$$\int_{\mathbf{I}} f\left(\mathbf{\Psi}(\mathbf{x})\right) \left|\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})\right| d\mathbf{x} = \sum_{\mathbf{J} \in \mathcal{J}_\mathbf{P}} \int_{\mathbf{J}} f\left(\mathbf{\Psi}(\mathbf{x})\right) \left|\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})\right| d\mathbf{x}.$$

Each $\mathbf{J} \in \mathcal{J}_\mathbf{P}$ is a closed rectangle. Hence, it is compact and path-connected. Since $\mathbf{\Psi} : \mathbf{I} \to \mathbb{R}^n$ is continuous, $\mathbf{\Psi}(\mathbf{I})$ is also compact and path-conneted. By the generalized mean value theorem, for each $\mathbf{J} \in \mathcal{J}_\mathbf{P}$, there exist $\mathbf{u}_\mathbf{J}$ and $\mathbf{v}_\mathbf{J}$ in $\mathbf{J}$ such that

$$\int_{\mathbf{\Psi}(\mathbf{J})} f(\mathbf{x}) d\mathbf{x} = f(\mathbf{\Psi}(\mathbf{u}_\mathbf{J}))\, \text{vol}\, (\mathbf{\Psi}(\mathbf{J}))$$

and

$$\int_{\mathbf{J}} f\left(\mathbf{\Psi}(\mathbf{x})\right) \left|\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})\right| d\mathbf{x} = f(\mathbf{\Psi}(\mathbf{v}_\mathbf{J})) \int_{\mathbf{J}} \left|\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})\right| d\mathbf{x}.$$

By the change of volume theorem,

$$\int_{\mathbf{J}} \left|\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})\right| d\mathbf{x} = \text{vol}\, (\mathbf{\Psi}(\mathbf{J})).$$

Since $\|\mathbf{u}_\mathbf{J} - \mathbf{v}_\mathbf{J}\| \leq \text{diam}\, \mathbf{J} < \delta$, we find that

$$|f(\mathbf{\Psi}(\mathbf{u}_\mathbf{J})) - f(\mathbf{\Psi}(\mathbf{v}_\mathbf{J}))| < \frac{\varepsilon}{\text{vol}\,(\mathbf{\Psi}(\mathbf{I}))}.$$

It follows that

$$\begin{aligned}
&\left|\int_{\mathbf{\Psi}(\mathbf{I})} f(\mathbf{x})d\mathbf{x} - \int_{\mathbf{I}} f\left(\mathbf{\Psi}(\mathbf{x})\right)\left|\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})\right| d\mathbf{x}\right| \\
&\leq \sum_{\mathbf{J}\in\mathcal{J}_{\mathbf{P}}} \left|\int_{\mathbf{\Psi}(\mathbf{J})} f(\mathbf{x})d\mathbf{x} - \int_{\mathbf{J}} f\left(\mathbf{\Psi}(\mathbf{x})\right)\left|\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})\right| d\mathbf{x}\right| \\
&= \sum_{\mathbf{J}\in\mathcal{J}_{\mathbf{P}}} \left|f(\mathbf{\Psi}(\mathbf{u}_{\mathbf{J}})) - f(\mathbf{\Psi}(\mathbf{v}_{\mathbf{J}}))\right| \text{ vol } \left(\mathbf{\Psi}(\mathbf{J})\right) \\
&< \frac{\varepsilon}{\text{vol}\left(\mathbf{\Psi}(\mathbf{I})\right)} \sum_{\mathbf{J}\in\mathcal{J}_{\mathbf{P}}} \text{vol } \left(\mathbf{\Psi}(\mathbf{J})\right) = \frac{\varepsilon}{\text{vol}\left(\mathbf{\Psi}(\mathbf{I})\right)} \times \text{vol}\left(\mathbf{\Psi}(\mathbf{I})\right) = \varepsilon.
\end{aligned}$$

This completes the proof.

Finally, we conclude the general case.

Conclusion of the Proof of the Change of Variables Theorem

As in the special case, it is sufficient to show that for any $\varepsilon > 0$,

$$\left|\int_{\mathbf{\Psi}(\mathfrak{D})} f(\mathbf{x})d\mathbf{x} - \int_{\mathfrak{D}} f\left(\mathbf{\Psi}(\mathbf{x})\right)\left|\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})\right| d\mathbf{x}\right| < \varepsilon.$$

Let us first proceed as in the proof of change of volume theorem. There is a positive number d and a compact set C such that $\overline{\mathfrak{D}} \subset C \subset \mathcal{O}$, and any point in $\mathbb{R}^n$ that has a distance less than d from a point in $\overline{\mathfrak{D}}$ lies in C. On the compact set C, the function $g : C \to \mathbb{R}$, $g(\mathbf{x}) = |\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})|$ is continuous. Therefore, it is bounded. Namely, there is a positive number M_g so that

$$|\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})| \leq M_g \qquad \text{for all } \mathbf{x} \in C.$$

Since we assume that the function $f : \mathbf{\Psi}(\mathfrak{D}) \to \mathbb{R}$ is bounded, there is a positive number M_f such that

$$|f(\mathbf{\Psi}(\mathbf{x}))| \leq M_f \qquad \text{for all } \mathbf{x} \in \mathfrak{D}.$$

Let $M = M_f M_g$, and let $\mathbf{I}$ be a rectangle that contains $\mathfrak{D}$. Given $\varepsilon > 0$, let $\mathbf{P}$ be a partition of $\mathbf{I}$ such that $|\mathbf{P}| < d$ and

$$L(\chi_{\mathfrak{D}}, \mathbf{P}) > \text{vol}\,(\mathfrak{D}) - \frac{\varepsilon}{2M}.$$

Let

$$\mathscr{A} = \{\mathbf{J} \in \mathcal{J}_{\mathbf{P}} \,|\, \mathbf{J} \cap \mathfrak{D} \neq \emptyset\}, \qquad \mathscr{B} = \{\mathbf{J} \in \mathcal{J}_{\mathbf{P}} \,|\, \mathbf{J} \subset \mathfrak{D}\}.$$

Since $|\mathbf{P}| < d$, each $\mathbf{J}$ in $\mathscr{A}$ is contained in C. Moreover,

$$L(\chi_{\mathfrak{D}}, \mathbf{P}) = \sum_{\mathbf{J} \in \mathscr{B}} \text{vol}\,(\mathbf{J}),$$

Denote by $\mathcal{Q}$ the set

$$\mathcal{Q} = \bigcup_{\mathbf{J} \in \mathscr{B}} \mathbf{J}.$$

Then $\mathcal{Q}$ is a compact subset of $\mathfrak{D}$, $\mathcal{S} = \mathfrak{D} \setminus \mathcal{Q}$ is Jordan measurable, and

$$\text{vol}(\mathfrak{D} \setminus \mathcal{Q}) = \text{vol}\,(\mathfrak{D}) - \sum_{\mathbf{J} \in \mathscr{B}} \text{vol}\,(\mathbf{J}) < \frac{\varepsilon}{2M}.$$

By additivitity theorem,

$$\int_{\mathbf{\Psi}(\mathfrak{D})} f(\mathbf{x}) d\mathbf{x} = \sum_{\mathbf{J} \in \mathscr{B}} \int_{\mathbf{\Psi}(\mathbf{J})} f(\mathbf{x}) d\mathbf{x} + \int_{\mathbf{\Psi}(\mathfrak{D} \setminus \mathcal{Q})} f(\mathbf{x}) d\mathbf{x},$$

$$\begin{aligned}\int_{\mathfrak{D}} f\left(\mathbf{\Psi}(\mathbf{x})\right) \left|\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})\right| d\mathbf{x} = &\sum_{\mathbf{J} \in \mathscr{B}} \int_{\mathbf{J}} f\left(\mathbf{\Psi}(\mathbf{x})\right) \left|\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})\right| d\mathbf{x} \\ &+ \int_{\mathfrak{D} \setminus \mathcal{Q}} f\left(\mathbf{\Psi}(\mathbf{x})\right) \left|\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})\right| d\mathbf{x}.\end{aligned}$$

Theorem 6.77 says that for each $\mathbf{J}$ in $\mathscr{B}$,

$$\int_{\mathbf{\Psi}(\mathbf{J})} f(\mathbf{x}) d\mathbf{x} = \int_{\mathbf{J}} f\left(\mathbf{\Psi}(\mathbf{x})\right) \left|\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})\right| d\mathbf{x}.$$

Therefore,

$$\left|\int_{\mathbf{\Psi}(\mathfrak{D})} f(\mathbf{x})d\mathbf{x} - \int_{\mathfrak{D}} f\left(\mathbf{\Psi}(\mathbf{x})\right)\left|\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})\right| d\mathbf{x}\right|$$
$$\leq \left|\int_{\mathbf{\Psi}(\mathfrak{D}\setminus\mathcal{Q})} f(\mathbf{x})d\mathbf{x}\right| + \left|\int_{\mathfrak{D}\setminus\mathcal{Q}} f\left(\mathbf{\Psi}(\mathbf{x})\right)\left|\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})\right| d\mathbf{x}\right|.$$

For the term $\int_{\mathbf{\Psi}(\mathfrak{D}\setminus\mathcal{Q})} f(\mathbf{x})d\mathbf{x}$, we have

$$\left|\int_{\mathbf{\Psi}(\mathfrak{D}\setminus\mathcal{Q})} f(\mathbf{x})d\mathbf{x}\right| \leq M_f \,\mathrm{vol}\left(\mathbf{\Psi}(\mathfrak{D}\setminus\mathcal{Q})\right).$$

By the change of volume theorem,

$$\mathrm{vol}\left(\mathbf{\Psi}(\mathfrak{D}\setminus\mathcal{Q})\right) = \int_{\mathfrak{D}\setminus\mathcal{Q}} \left|\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})\right| d\mathbf{x} \leq M_g \mathrm{vol}\left(\mathfrak{D}\setminus\mathcal{Q}\right).$$

Therefore,

$$\left|\int_{\mathbf{\Psi}(\mathfrak{D}\setminus\mathcal{Q})} f(\mathbf{x})d\mathbf{x}\right| \leq M_f M_g \mathrm{vol}\left(\mathfrak{D}\setminus\mathcal{Q}\right) = M\mathrm{vol}\left(\mathfrak{D}\setminus\mathcal{Q}\right) < \frac{\varepsilon}{2}.$$

Similarly,

$$\left|\int_{\mathfrak{D}\setminus\mathcal{Q}} f\left(\mathbf{\Psi}(\mathbf{x})\right)\left|\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})\right| d\mathbf{x}\right| \leq M\mathrm{vol}\left(\mathfrak{D}\setminus\mathcal{Q}\right) < \frac{\varepsilon}{2}.$$

This gives

$$\left|\int_{\mathbf{\Psi}(\mathfrak{D})} f(\mathbf{x})d\mathbf{x} - \int_{\mathfrak{D}} f\left(\mathbf{\Psi}(\mathbf{x})\right)\left|\det \mathbf{D}\mathbf{\Psi}(\mathbf{x})\right| d\mathbf{x}\right| < \varepsilon,$$

which completes the proof.

6.7 Some Important Integrals and Their Applications

Up to now we have only discussed multiple integrals for bounded functions $f : \mathfrak{D} \to \mathbb{R}$ defined on bounded domains. For practical applications, we need to consider improper integrals where the function is not bounded or the domain is not bounded. As in the single variable case, we need to take limits. In the multi-variable case, things become considerably more complicated. Interested readers can read the corresponding sections in the book [Zor16]. In this section, we use theories learned in multiple integrals to derive some explicit formulas of improper integrals of single-variable functions, without introducing the definition of improper multiple integrals. We then give some applications of these formulas.

Proposition 6.78

For any positive number a,

$$\int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}}.$$

Proof

Since the function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^{-ax^2}$ is positive for all $x \in \mathbb{R}$,

$$\int_{-\infty}^{\infty} e^{-ax^2} dx = \lim_{L\to\infty} \int_{-L}^{L} e^{-ax^2} dx.$$

Given a positive number R, we consider the double integral

$$I_R = \int_{\overline{B(\mathbf{0},R)}} e^{-a(x^2+y^2)} dxdy.$$

For any positive number L,

$$\overline{B(\mathbf{0}, L)} \subset [-L, L] \times [-L, L] \subset \overline{B(\mathbf{0}, \sqrt{2}L)}.$$

Since the function $g : \mathbb{R}^2 \to \mathbb{R}$, $g(x) = e^{-a(x^2+y^2)}$ is positive,

$$I_L \leq \int_{[-L,L]\times[-L,L]} e^{-a(x^2+y^2)} dxdy \leq I_{\sqrt{2}L}. \tag{6.19}$$

Using polar coordinates, we find that

$$I_R = \int_0^{2\pi} \int_0^R e^{-ar^2} r dr d\theta = 2\pi \left[-\frac{e^{-ar^2}}{2a} \right]_0^R = \frac{\pi}{a} \left(1 - e^{-aR^2} \right).$$

Thus,

$$\lim_{R\to\infty} I_R = \frac{\pi}{a}.$$

Eq. (6.19) then implies that

$$\lim_{L\to\infty} \int_{[-L,L]\times[-L,L]} e^{-a(x^2+y^2)} dxdy = \frac{\pi}{a}.$$

By Fubini's theorem,

$$\int_{[-L,L]\times[-L,L]} e^{-a(x^2+y^2)} dxdy = \left(\int_{-L}^{L} e^{-x^2} dx \right)^2.$$

Thus, we conclude that

$$\int_{-\infty}^{\infty} e^{-ax^2} dx = \lim_{L\to\infty} \int_{-L}^{L} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}}.$$

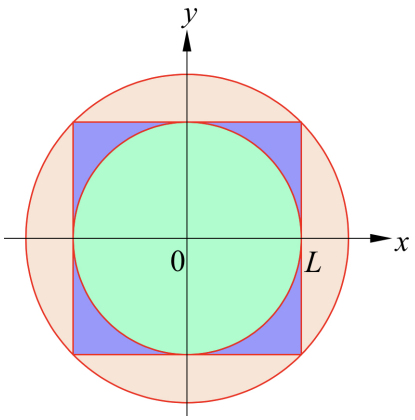


Figure 6.48: $\overline{B(\mathbf{0}, L)} \subset [-L, L] \times [-L, L] \subset \overline{B(\mathbf{0}, \sqrt{2}L)}$.

The improper integral $\int_{-\infty}^{\infty} e^{-ax^2} dx$ with $a > 0$ plays an important role in various areas of mathematics. For example, in probability theorem, the probability

density function of a normal random variable with mean μ and standard deviation σ is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

The normalization factor $1/(\sqrt{2\pi}\sigma)$ is required such that

$$\int_{-\infty}^{\infty} f(x)dx = 1,$$

which ensures that total probability is 1.

Recall that we have defined the gamma function $\Gamma(s)$ for a real number $s > 0$ by the improper integral

$$\Gamma(s) = \int_0^{\infty} t^{s-1}e^{-t}dt.$$

The value of $\Gamma(1)$ is easy to compute.

$$\Gamma(1) = \int_0^{\infty} e^{-t}dt = 1.$$

Using integration by parts, one can show that

$$\Gamma(s+1) = s\Gamma(s) \qquad \text{when } s > 0. \tag{6.20}$$

From this, we find that

$$\Gamma(n+1) = n! \qquad \text{for all } n \in \mathbb{Z}^+.$$

The value of $\Gamma(s)$ when $s = 1/2$ is also of particular interest.

Theorem 6.79

The value of the gamma function $\Gamma(s)$ at $s = 1/2$ is

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}.$$

Proof

By Proposition 6.78,

$$\sqrt{\pi} = 2\int_0^{\infty} e^{-x^2}dx = 2\lim_{a\to 0^+}\lim_{L\to\infty}\int_a^L e^{-x^2}dx.$$

Making a change of variables $t = x^2$, we find that

$$2\int_a^L e^{-x^2}dx = \int_{a^2}^{L^2} t^{-\frac{1}{2}}e^{-t}dt.$$

Therefore,

$$\begin{aligned}\Gamma\left(\frac{1}{2}\right) &= \int_0^\infty t^{-\frac{1}{2}}e^{-t}dt = \lim_{a\to 0^+}\lim_{L\to\infty}\int_{a^2}^{L^2} t^{-\frac{1}{2}}e^{-t}dt \\ &= 2\lim_{a\to 0^+}\int_a^L e^{-x^2}dx = \sqrt{\pi}.\end{aligned}$$

Another useful formula we have mentioned in volume I is the formula for the beta function $B(\alpha,\beta)$ defined as

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt \qquad \text{when } \alpha>0, \beta>0.$$

It is easy to show that the integral is indeed convergent when α and β are positive. We have the following recursive formula.

Lemma 6.80

For $\alpha > 0$ and $\beta > 0$, we have

$$B(\alpha,\beta) = \frac{(\alpha+\beta+1)(\alpha+\beta)}{\alpha\beta}B(\alpha+1,\beta+1).$$

Proof

First notice that for $\alpha > 0$ and $\beta > 0$,

$$\begin{aligned}\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt &= \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}(1-t+t)dt \\ &= \int_0^1 t^{\alpha-1}(1-t)^{\beta}dt + \int_0^1 t^{\alpha}(1-t)^{\beta-1}dt.\end{aligned}$$

This gives

$$B(\alpha,\beta) = B(\alpha+1,\beta) + B(\alpha,\beta+1).$$

Apply this formula again to the two terms on the right, we find that

$$B(\alpha,\beta) = B(\alpha+2,\beta) + 2B(\alpha+1,\beta+1) + B(\alpha,\beta+2).$$

Using integration by parts, one can show that when $\alpha > 0$ and $\beta > 0$,

$$\int_0^1 t^\alpha (1-t)^{\beta-1} dt = \frac{\alpha}{\beta}\int_0^1 t^{\alpha-1}(1-t)^\beta dt.$$

This gives

$$B(\alpha+1,\beta) = \frac{\alpha}{\beta}B(\alpha,\beta+1),$$

Therefore,

$$\begin{aligned} B(\alpha,\beta) &= \left(\frac{\alpha+1}{\beta} + 2 + \frac{\beta+1}{\alpha}\right) B(\alpha+1,\beta+1) \\ &= \frac{(\alpha+\beta+1)(\alpha+\beta)}{\alpha\beta} B(\alpha+1,\beta+1). \end{aligned}$$

Now we can derive the explicit formula for the beta function.

Theorem 6.81 The Beta Function

For any positive real numbers α and β,

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Proof

We first consider the case where $\alpha > 1$ and $\beta > 1$. Let $g : [0,\infty) \times [0,\infty) \to \mathbb{R}$ be the function defined as

$$g(u,v) = u^{\alpha-1}v^{\beta-1}e^{-u-v}.$$

This is a continuous function. For $L > 0$, let

$$\begin{aligned} \mathcal{U}_L &= \{(t,w) \,|\, 0 \le t \le 1, 0 \le w \le L\}, \\ \mathfrak{D}_L &= \{(u,v) \,|\, u \ge 0, v \ge 0, u+v \le L\}. \end{aligned}$$

Consider the map $\mathbf{\Psi} : \mathbb{R}^2 \to \mathbb{R}^2$ defined as

$$(u, v) = \mathbf{\Psi}(t, w) = (tw, (1 - t)w).$$

Notice that $\mathbf{\Psi}$ maps the interior of $\mathcal{U}_L$ one-to-one onto the interior of $\mathfrak{D}_L$. The Jacobian of this map is

$$\frac{\partial(u, v)}{\partial(t, w)} = \det \begin{bmatrix} w & t \\ -w & 1 - t \end{bmatrix} = w.$$

Thus, $\mathbf{\Psi} : (0, 1) \times (0, \infty) \to \mathbb{R}^2$ is a smooth change of variables. By taking limits, the change of variables theorem implies that

$$\int_{\mathcal{U}_L} (g \circ \mathbf{\Psi})(t, w) \frac{\partial(u, v)}{\partial(t, w)} dt dw = \int_{\mathfrak{D}_L} g(u, v) du dv. \tag{6.21}$$

Now Fubini's theorem says that

$$\int_{\mathcal{U}_L} (g \circ \mathbf{\Psi})(t, w) \frac{\partial(u, v)}{\partial(t, w)} dt dw = \int_0^1 t^{\alpha-1}(1 - t)^{\beta-1} dt \int_0^L w^{\alpha+\beta-1} e^{-w} dw.$$

Thus,

$$\lim_{L\to\infty} \int_{\mathcal{U}_L} (g \circ \mathbf{\Psi})(t, w) \frac{\partial(u, v)}{\partial(t, w)} dt dw = \Gamma(\alpha + \beta) B(\alpha, \beta).$$

On the other hand, we notice that

$$\left[0, \frac{L}{2}\right]^2 \subset \mathfrak{D}_L \subset [0, L]^2.$$

Since the function $g : [0, \infty) \times [0, \infty) \to \mathbb{R}$ is nonnegative, this implies that

$$I_{\frac{L}{2}} \leq \int_{\mathfrak{D}_L} g(u, v) du dv \leq I_L, \tag{6.22}$$

where

$$I_L = \int_{[0,L]^2} g(u, v) du dv = \int_0^L u^{\alpha-1} e^{-u} du \int_0^L v^{\beta-1} e^{-v} dv.$$

By the definition of the gamma function,

$$\lim_{L\to\infty} I_L = \Gamma(\alpha)\Gamma(\beta).$$

Eq. (6.22) then implies that

$$\lim_{L\to\infty} \int_{\mathfrak{D}_L} g(u,v)dudv = \Gamma(\alpha)\Gamma(\beta).$$

It follows from (6.21) that

$$\Gamma(\alpha+\beta)B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta),$$

which gives the desired formula when $\alpha > 1$ and $\beta > 1$.
For the general case where $\alpha > 0$ and $\beta > 0$, Lemma 6.80 and (6.20) give

$$B(\alpha,\beta) = \frac{(\alpha+\beta+1)(\alpha+\beta)}{\alpha\beta}\frac{\Gamma(\alpha+1)\Gamma(\beta+1)}{\Gamma(\alpha+\beta+2)} = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

This completes the proof.

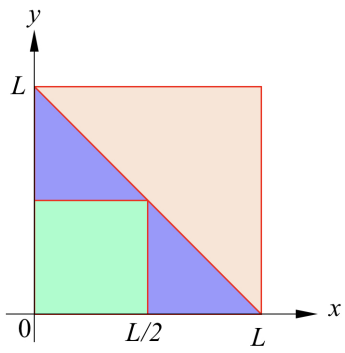


Figure 6.49: $\left[0,\frac{L}{2}\right]^2 \subset \{(u,v)\,|\,u\geq 0, v\geq 0, u+v\leq L\} \subset [0,L]^2$.

Now we give an interesting application of the formula of the beta function.

Theorem 6.82

For $n \geq 1$, the volume of the n-ball of radius a,

$$B_n(a) = \left\{(x_1, \ldots, x_n) \in \mathbb{R}^n \,|\, x_1^2 + \cdots + x_n^2 \leq a^2\right\},$$

is equal to

$$V_n(a) = \frac{\pi^{\frac{n}{2}} a^n}{\Gamma\left(\frac{n+2}{2}\right)}.$$

Proof

It is easy to see that for any $a > 0$,

$$V_n(a) = V_n a^n, \qquad \text{where } V_n = V_n(1).$$

Since $B_1(1) = [-1, 1]$, we find that $V_1 = 2$. For $n \geq 2$, notice that for fixed $-1 \leq y \leq 1$, the ball $B_n(1)$ intersects the plane $x_n = y$ on the set

$$S_n(y) = \left\{(x_1, \ldots, x_{n-1}, y) \,|\, x_1^2 + \ldots + x_{n-1}^2 \leq 1 - y^2\right\}.$$

Fubini's theorem implies that

$$\begin{aligned} V_n &= \int_{-1}^{1} \int_{B_{n-1}\left(\sqrt{1-y^2}\right)} dx_1 \ldots dx_{n-1} dy = \int_{-1}^{1} (1-y^2)^{\frac{n-1}{2}} V_{n-1} dy \\ &= 2V_{n-1} \int_0^1 (1-y^2)^{\frac{n-1}{2}} dy = V_{n-1} \int_0^1 t^{-\frac{1}{2}} (1-t)^{\frac{n-1}{2}} dt \\ &= V_{n-1} B\left(\frac{1}{2}, \frac{n+1}{2}\right) = \sqrt{\pi} V_{n-1} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n+2}{2}\right)}. \end{aligned}$$

This formula is still correct when $n = 1$ if we define $V_0 = 1$. Therefore,

$$V_n = \prod_{k=1}^{n} \frac{V_k}{V_{k-1}} = \prod_{k=1}^{n} \left(\sqrt{\pi} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k+2}{2}\right)}\right) = \frac{\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n+2}{2}\right)}.$$

Chapter 7

Fourier Series and Fourier Transforms

In this chapter, we shift our attention to the theory of Fourier series and Fourier transforms. In volume I, we have considered expansions of functions as power series, which are limits of polynomials. In this chapter, we consider expansions of functions in another class of infinitely differentiable functions – the trigonometric functions $\sin x$ and $\cos x$. The reason to consider $\sin x$ and $\cos x$ is that they are representative of periodic functions.

Recall that a function $f : \mathbb{R} \to \mathbb{R}$ is said to be periodic if there is a positive number p so that

$$f(x + p) = f(x) \qquad \text{for all } x \in \mathbb{R}.$$

Such a number p is called a period of the function f. If p is a period of f, then for any positive integer n, np is also a period of f.

The functions $\sin x$ and $\cos x$ are periodic functions of period 2π. If $f : \mathbb{R} \to \mathbb{R}$ is a periodic function of period $p = 2L$, then the function $g : \mathbb{R} \to \mathbb{R}$ defined as

$$g(x) = f\left(\frac{\pi x}{L}\right)$$

is periodic of period 2π. Hence, we can concentrate on functions that are periodic of period 2π.

The celebrated Euler formula

$$e^{ix} = \cos x + i \sin x$$

connects the trigonometric functions $\sin x$, $\cos x$ with the exponential function with imaginary arguments. Hence, in this chapter, we shift our paradigm and consider complex-valued functions $f : D \to \mathbb{C}$ defined on a subset D of $\mathbb{R}$. Since a complex number $z = x + iy$ with real part x and imaginary part y can be identified with the point (x, y) in $\mathbb{R}^2$, such a function can be regarded as a function $f : D \to \mathbb{R}^2$, so that derivative and integrals are defined componentwise. More

precisely, given $x \in D$, we write $f(x) = u(x) + iv(x)$, where $u(x)$ and $v(x)$ are respectively the real part and imaginary part of $f(x)$. If x_0 is an interior point of D, we say that f is differentiable at x_0 if the limit

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

exists. This is if and only if both $u : D \to \mathbb{R}$ and $v : D \to \mathbb{R}$ are differentiable at x_0, and we have

$$f'(x_0) = u'(x_0) + iv'(x_0).$$

Similarly, if $[a, b]$ is a closed interval that is contained in D, we say that f is Riemann integrable over $[a, b]$ if and only if both u and v are Riemann integrable over $[a, b]$, and we have

$$\int_a^b f(x)dx = \int_a^b u(x)dx + i \int_a^b v(x)dx.$$

If $F : [a, b] \to \mathbb{C}$ is a continuously differentiable function, the fundamental theorem of calculus implies that

$$F(b) - F(a) = \int_a^b F'(x)dx.$$

7.1 Orthogonal Systems of Functions and Fourier Series

In the following, let $I = [a, b]$ be a compact interval in $\mathbb{R}$ unless otherwise specified. Denote by $\mathcal{R}(I, \mathbb{C})$ the set of all complex-valued functions $f : I \to \mathbb{C}$ that are Riemann integrable. Given two functions f and g in $\mathcal{R}(I, \mathbb{C})$, their sum $f + g$ is the function $(f + g) : I \to \mathbb{C}$,

$$(f + g)(x) = f(x) + g(x).$$

If α is a complex number, the scalar product of α with f is the function $(\alpha f) : I \to \mathbb{C}$, where

$$(\alpha f)(x) = \alpha f(x).$$

With the addition and scalar multiplication thus defined, $\mathcal{R}(I, \mathbb{C})$ is a complex vector space. From the theory of integration, we know that the set of complex-valued continuous functions on I, denoted by $C(I, \mathbb{C})$, is a subspace of $\mathcal{R}(I, \mathbb{C})$.

If $f : I \to \mathbb{C}$ is Riemann integrable, so does its complex conjugate $\overline{f} : I \to \mathbb{C}$ defined as

$$\overline{f}(x) = \overline{f(x)}.$$

In volume I, we have proved that if two real-valued functions $f : I \to \mathbb{R}$ and $g : I \to \mathbb{R}$ are Riemann integrable, so is their product $(fg) : I \to \mathbb{R}$. Using this, it is easy to check that if $f : I \to \mathbb{C}$ and $g : I \to \mathbb{C}$ are Riemann integrable complex-valued functions, $(f\overline{g}) : I \to \mathbb{C}$ is also Riemann integrable.

Proposition 7.1

Given f and g in $\mathcal{R}(I, \mathbb{C})$, define

$$\langle f, g\rangle = \int_a^b f(x)\overline{g(x)}dx.$$

For any f, g, h in $\mathcal{R}(I, \mathbb{C})$, and any complex numbers α and β, we have the followings.

(a) $\langle g, f\rangle = \overline{\langle f, g\rangle}$.

(b) $\langle \alpha f + \beta g, h\rangle = \alpha\langle f, h\rangle + \beta\langle g, h\rangle$.

(c) $\langle f, f\rangle \geq 0$.

We call $\langle \cdot\ ,\ \cdot \rangle$ a positive *semi-definite* inner product on $\mathcal{R}(I, \mathbb{C})$.

It follows from (a) and (b) that

$$\langle f, \alpha g + \beta h\rangle = \overline{\alpha}\langle f, g\rangle + \overline{\beta}\langle f, h\rangle.$$

More generally, we have

$$\left\langle \sum_{i=1}^m \alpha_i f_i,\ \sum_{j=1}^n \beta_j g_j \right\rangle = \sum_{i=1}^m \sum_{j=1}^n \alpha_i \overline{\beta_j}\langle f_i, g_j\rangle. \tag{7.1}$$

If $f(x) = u(x) + iv(x)$, where $u(x) = \text{Re}\, f(x)$ and $v(x) = \text{Im}\, f(x)$, then

$$\langle f, f\rangle = \int_a^b \left(u(x)^2 + v(x)^2\right) dx.$$

Notice that $\langle f, f\rangle = 0$ does not imply that $f = 0$. For example, take any c in $[a, b]$, and define the function $f : I \to \mathbb{C}$ by

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

Then

$$\langle f, f\rangle = \int_a^b f(x)\overline{f(x)}dx = 0$$

even though f is not the zero function. This is why we call $\langle \cdot\ ,\ \cdot \rangle$ a positive semi-definite inner product. Restricted to the subspace of continuous functions $C(I, \mathbb{C})$, $\langle \cdot\ ,\ \cdot \rangle$ is a positive definite inner product, or simply an inner product in the usual sense.

Using the positive semi-definite inner product, we can define a semi-norm on $\mathcal{R}(I, \mathbb{C})$.

Definition 7.1 The L^2 Semi-Norm

Given $f : I \to \mathbb{C}$ in $\mathcal{R}(I, \mathbb{C})$, the semi-norm of f is defined as

$$\|f\| = \sqrt{\langle f, f\rangle} = \sqrt{\int_a^b |f(x)|^2 dx}.$$

It has the following properties.

Proposition 7.2

Given f in $\mathcal{R}(I, \mathbb{C})$ and $\alpha \in \mathbb{C}$, we have the followings.

(a) $\|f\| \geq 0$.

(b) $\|\alpha f\| = |\alpha|\|f\|$.

The Cauchy-Schwarz inequality still holds for the positive semi-definite inner product on $\mathcal{R}(I, \mathbb{C})$.

Proposition 7.3 Cauchy-Schwarz Inequality

Given f and g in $\mathcal{R}(I, \mathbb{C})$,

$$|\langle f, g\rangle| \leq \|f\|\|g\|.$$

The proof is exactly the same as for an inner product on a real vector space. An immediate consequence of the Cauchy-Schwarz inequality is the triangle inequality.

Proposition 7.4 Triangle Inequality

Let $f_1, \ldots, f_n$ be functions in $\mathcal{R}(I, \mathbb{C})$ and let $\alpha_1, \ldots, \alpha_n$ be complex numbers. We have

$$\|\alpha_1 f_1 + \cdots + \alpha_n f_n\| \leq |\alpha_1|\|f_1\| + \cdots + |\alpha_n|\|f_n\|.$$

The proof is also the same as for a real inner product. One consider the case $n = 2$ first and then prove the general case by induction on n.

We can define orthogonality on $\mathcal{R}(I, \mathbb{C})$ the same way as for a real inner product space.

Definition 7.2 Orthogonality

Given two functions f and g in $\mathcal{R}(I, \mathbb{C})$, we say that they are orthogonal if $\langle f, g\rangle = 0$.

Example 7.1

Let $I = [0, 2\pi]$. For $n \in \mathbb{Z}$, define $\phi_n : I \to \mathbb{C}$ by $\phi_n(x) = e^{inx}$. Show that if m and n are distinct integers, then ϕ_m and ϕ_n are orthogonal.

Solution

Notice that

$$\langle \phi_m, \phi_n\rangle = \int_0^{2\pi} \phi_m(x)\overline{\phi_n(x)}dx = \int_0^{2\pi} e^{i(m-n)x}dx.$$

Since $m \neq n$, and

$$\frac{d}{dx}e^{i(m-n)x} = i(m-n)e^{i(m-n)x},$$

fundamental theorem of calculus implies that

$$\langle \phi_m, \phi_n \rangle = \left[\frac{e^{i(m-n)x}}{i(m-n)} \right]_0^{2\pi} = 0.$$

Hence, ϕ_m and ϕ_n are orthogonal.

Definition 7.3 Orthogonal System and Orthonormal System

Let $\mathcal{S} = \{\phi_\alpha \,|\, \alpha \in J\}$ be a subset of functions in $\mathcal{R}(I, \mathbb{C})$ indexed by the set J. We say that $\mathcal{S}$ is an orthogonal system of functions if

$$\langle \phi_\alpha, \phi_\beta \rangle = 0 \qquad \text{whenever } \alpha \neq \beta,$$

and

$$\|\phi_\alpha\| \neq 0 \qquad \text{for all } \alpha \in J.$$

We say that $\mathcal{S}$ is an orthonormal system of functions if it is an orthogonal system and

$$\|\phi_\alpha\| = 1 \qquad \text{for all } \alpha \in J.$$

Notice that in our definition of orthogonal system, we have an additional condition that each element in the set $\mathcal{S}$ cannot have zero norm. By definition, it is obvious that if $\mathcal{S}$ is an orthogonal system, then any subset of $\mathcal{S}$ is also an orthogonal system. The same holds for orthonormal systems.

Example 7.2

Let $I = [0, 2\pi]$. For $n \in \mathbb{Z}$, define $\phi_n : I \to \mathbb{C}$ by $\phi_n(x) = e^{inx}$. Then

$$\|\phi_n\|^2 = \int_0^{2\pi} e^{inx}e^{-inx}dx = 2\pi.$$

Example 7.1 implies that $\mathcal{S} = \{\phi_n \,|\, n \in \mathbb{Z}\}$ is an orthogonal system.

If we let $\varphi_n : I \to \mathbb{R}$, $n \in \mathbb{Z}$ be the function

$$\varphi_n(x) = \frac{\phi_n(x)}{\|\phi_n\|} = \frac{e^{inx}}{\sqrt{2\pi}},$$

then $\widetilde{\mathcal{S}} = \{\varphi_n \,|\, n \in \mathbb{Z}\}$ is an orthonormal system.

Using the semi-norm, we can define a relation $\sim$ on $\mathcal{R}(I, \mathbb{C})$ in the following way. We say that $f \sim g$ if and only if $\|f - g\| = 0$. It is easy to check that this is an equivalence relation. Reflexivity and symmetry are obvious. For transitivity, we note that if $f \sim g$ and $g \sim h$, then $\|f - g\| = 0$ and $\|g - h\| = 0$. It follows from triangle inequality that

$$\|f - h\| \leq \|f - g\| + \|g - h\| = 0.$$

This implies that $\|f - h\| = 0$, and thus $f \sim h$. Hence, $\sim$ is an equivalence relation on $\mathcal{R}(I, \mathbb{C})$, which we call L^2-equivalent.

Definition 7.4 L^2 Equivalent Functions

Two Riemann integrable functions $f : I \to \mathbb{C}$ and $g : I \to \mathbb{C}$ are L^2-equivalent if

$$\|f - g\| = 0.$$

Example 7.3

Let $I = [a, b]$, and let S be a finite subset of I. If $f : I \to \mathbb{C}$ and $g : I \to \mathbb{C}$ are two Riemann integrable functions and

$$f(x) = g(x) \qquad \text{for all } x \in [a, b] \setminus S,$$

then f and g are L^2-equivalent.

Regarding $\mathcal{R}(I, \mathbb{C})$ as an additive group, the subset $\mathcal{K}(I, \mathbb{C})$ that contains all the functions in $\mathcal{R}(I, \mathbb{C})$ that have zero norm is a normal subgroup. They are functions that are L^2-equivalent to the zero function. Denote by

$$\widehat{\mathcal{R}}(I, \mathbb{C}) = \mathcal{R}(I, \mathbb{C})/\mathcal{K}(I, \mathbb{C}) = \mathcal{R}(I, \mathbb{C})/\sim$$

the quotient group. Then each element of $\widehat{\mathcal{R}}(I,\mathbb{C})$ is an L^2 equivalent class of functions.

If u is in $\mathcal{K}(I,\mathbb{C})$, g is in $\mathcal{R}(I,\mathbb{C})$, then Cauchy-Schwarz inequality implies that

$$|\langle u, g\rangle| \leq \|u\|\|g\| = 0.$$

Thus, $\langle u, g\rangle = 0$. If f is L^2 equivalent to f_1, g is L^2 equivalent to g_1, there exists u and v in $\mathcal{K}(I,\mathbb{C})$ such that $f_1 = f + u$ and $g_1 = g + v$. Therefore,

$$\langle f_1, g_1\rangle = \langle f + u, g + v\rangle = \langle f, g\rangle + \langle u, g\rangle + \langle f, v\rangle + \langle u, v\rangle = \langle f, g\rangle.$$

Hence, the positive semi-definite inner product $\langle\ \cdot\ ,\ \cdot\ \rangle$ on $\mathcal{R}(I,\mathbb{C})$ induces an infinite product on $\widehat{\mathcal{R}}(I,\mathbb{C})$ by

$$\langle [f], [g]\rangle = \langle f, g\rangle.$$

If $[f] \in \widehat{\mathcal{R}}(I,\mathbb{C})$ is such that

$$\langle [f], [f]\rangle = \langle f, f\rangle = 0,$$

then f is in $\mathcal{K}(I,\mathbb{C})$, and thus, $[f] = [0]$. This says that the infinite product $\langle\ \cdot\ ,\ \cdot\ \rangle$ on $\widehat{\mathcal{R}}(I,\mathbb{C})$ is positive definite. The additional condition we impose on a subset $\mathcal{S}$ of $\mathcal{R}(I,\mathbb{C})$ to be an orthogonal system just means that none of the elements in $\mathcal{S}$ is L^2-equivalent to the zero function.

For an orthogonal system, we have the following from (7.1).

Theorem 7.5 Generalized Pythagoras Theorem

Let $\mathcal{S} = \{\phi_k \,|\, 1 \leq k \leq n\}$ be an orthogonal system of functions in $\mathcal{R}(I,\mathbb{C})$. For any complex numbers $\alpha_1, \ldots, \alpha_n$,

$$\|\alpha_1\phi_1 + \cdots + \alpha_n\phi_n\|^2 = |\alpha_1|^2\|\phi_1\|^2 + \cdots + |\alpha_n|^2\|\phi_n\|^2.$$

The functions $\phi_n : [0, 2\pi] \to \mathbb{C}$, $f_n(x) = e^{inx}$, $n \in \mathbb{Z}$ are easy to deal with because of $\dfrac{d}{dx}e^{ax} = ae^{ax}$ for any complex numbers a. The drawback is they are complex-valued functions. Since

$$e^{inx} = \cos nx + i \sin nx,$$

if one wants to work with real-valued functions, one should consider the functions $\cos nx$ and $\sin nx$.

Proposition 7.6

Let $I = [0, 2\pi]$, and define the functions $C_n : I \to \mathbb{R}$, $n \geq 0$, and $S_n : I \to \mathbb{R}$, $n \geq 1$ by

$$C_n(x) = \cos nx, \qquad S_n(x) = \sin nx.$$

Then $\mathscr{B} = \{C_n \,|\, n \geq 0\} \cup \{S_n \,|\, n \geq 1\}$ is an orthogonal system, and $\|C_0\| = \sqrt{2\pi}$,

$$\|C_n\| = \|S_n\| = \sqrt{\pi} \qquad \text{when } n \geq 1.$$

Proof

For $n \in \mathbb{Z}$, let $\phi_n : [0, 2\pi] \to \mathbb{C}$ be the function $\phi_n(x) = e^{inx}$. Then $C_0 = \phi_0$, and when $n \in \mathbb{Z}^+$,

$$C_n = \frac{\phi_n + \phi_{-n}}{2}, \qquad S_n = \frac{\phi_n - \phi_{-n}}{2i}.$$

Since $\{\phi_n \,|\, n \in \mathbb{Z}\}$ is an orthogonal system, we find that for $n \in \mathbb{Z}^+$,

$$\langle C_0, C_n \rangle = \frac{1}{2}\langle \phi_0, \phi_n \rangle + \frac{1}{2}\langle \phi_0, \phi_{-n} \rangle = 0,$$

$$\langle C_0, S_n \rangle = \frac{i}{2}\langle \phi_0, \phi_n \rangle - \frac{i}{2}\langle \phi_0, \phi_{-n} \rangle = 0.$$

For $m, n \in \mathbb{Z}^+$ such that $m \neq n$,

$$\langle C_m, C_n \rangle = \frac{1}{4}\left(\langle \phi_m, \phi_n \rangle + \langle \phi_m, \phi_{-n} \rangle + \langle \phi_{-m}, \phi_n \rangle + \langle \phi_{-m}, \phi_{-n} \rangle\right) = 0,$$

$$\langle S_m, S_n \rangle = \frac{1}{4}\left(\langle \phi_m, \phi_n \rangle - \langle \phi_m, \phi_{-n} \rangle - \langle \phi_{-m}, \phi_n \rangle + \langle \phi_{-m}, \phi_{-n} \rangle\right) = 0.$$

For $m, n \in \mathbb{Z}^+$, considering the cases $m = n$ and $m \neq n$ separately, we find that

$$\langle S_m, C_n \rangle = \frac{1}{4i}\left(\langle \phi_m, \phi_n \rangle + \langle \phi_m, \phi_{-n} \rangle - \langle \phi_{-m}, \phi_n \rangle - \langle \phi_{-m}, \phi_{-n} \rangle\right) = 0.$$

These show that $\mathscr{B}$ is an orthogonal system. For $n \in \mathbb{Z}^+$, since $\|\phi_n\| = \|\phi_{-n}\| = \sqrt{2\pi}$, and ϕ_n and ϕ_{-n} are orthogonal, we have

$$\|C_n\|^2 = \left\| \frac{1}{2}\phi_n + \frac{1}{2}\phi_{-n} \right\|^2 = \frac{1}{4}\|\phi_n\|^2 + \frac{1}{4}\|\phi_{-n}\|^2 = \pi,$$

$$\|S_n\|^2 = \left\| \frac{1}{2i}\phi_n - \frac{1}{2i}\phi_{-n} \right\|^2 = \frac{1}{4}\|\phi_n\|^2 + \frac{1}{4}\|\phi_{-n}\|^2 = \pi.$$

These complete the proof.

Given a finite subset $\mathcal{S} = \{f_1, \ldots, f_n\}$ of $\mathcal{R}(I, \mathbb{C})$, let

$$W_{\mathcal{S}} = \operatorname{span} \mathcal{S} = \{c_1 f_1 + \cdots + c_n f_n \,|\, c_1, \ldots, c_n \in \mathbb{C}\}$$

be the subspace of $\mathcal{R}(I, \mathbb{C})$ spanned by $\mathcal{S}$. We say that an element g of $\mathcal{R}(I, \mathbb{C})$ is orthogonal to $W_{\mathcal{S}}$ if it is orthogonal to each $f \in W_{\mathcal{S}}$. This is if and only if g is orthogonal to f_k for all $1 \le k \le n$. The projection theorem says the following.

Theorem 7.7 Projection Theorem

Let $\mathcal{S} = \{\phi_1, \ldots, \phi_n\}$ be an orthogonal system of functions in $\mathcal{R}(I, \mathbb{C})$, and let $W_{\mathcal{S}}$ be the subspace of $\mathcal{R}(I, \mathbb{C})$ spanned by $\mathcal{S}$. Given f in $\mathcal{R}(I, \mathbb{C})$, there is a unique $g \in W_{\mathcal{S}}$ such that $f - g$ is orthogonal to $W_{\mathcal{S}}$. It is called the projection of the function f onto the subspace $W_{\mathcal{S}}$, denoted by $\operatorname{proj}_{W_{\mathcal{S}}} f$, and it is given by

$$\operatorname{proj}_{W_{\mathcal{S}}} f = \sum_{k=1}^{n} \frac{\langle f, \phi_k \rangle}{\langle \phi_k, \phi_k \rangle} \phi_k = \frac{\langle f, \phi_1 \rangle}{\langle \phi_1, \phi_1 \rangle} \phi_1 + \cdots + \frac{\langle f, \phi_n \rangle}{\langle \phi_n, \phi_n \rangle} \phi_n.$$

For any $h \in W_{\mathcal{S}}$,

$$\|f - h\| \ge \|f - \operatorname{proj}_{W_{\mathcal{S}}} f\|.$$

Proof

Assume that g is a function in $W_{\mathcal{S}}$ such that $f - g$ is orthogonal to $W_{\mathcal{S}}$. Then there exist complex numbers $\alpha_1, \ldots, \alpha_n$ such that

$$g = \alpha_1 \phi_1 + \cdots + \alpha_n \phi_n.$$

Since $\langle \phi_k, \phi_l \rangle = 0$ if $k \neq l$, we find that

$$\langle g, \phi_k \rangle = \alpha_k \langle \phi_k, \phi_k \rangle \qquad \text{for } 1 \leq k \leq n.$$

Since $f - g$ is orthogonal to $W_{\mathcal{S}}$, $\langle f - g, \phi_k \rangle = 0$ for all $1 \leq k \leq n$. This gives $\langle f, \phi_k \rangle = \langle g, \phi_k \rangle$, and thus

$$\alpha_k \langle \phi_k, \phi_k \rangle = \langle f, \phi_k \rangle \qquad \text{for } 1 \leq k \leq n.$$

Hence, we must have

$$\alpha_k = \frac{\langle f, \phi_k \rangle}{\langle \phi_k, \phi_k \rangle}.$$

This implies the uniqueness of g if it exists. It is easy to check that the function

$$g = \sum_{k=1}^{n} \frac{\langle f, \phi_k \rangle}{\langle \phi_k, \phi_k \rangle} \phi_k$$

is indeed a function in $W_{\mathcal{S}}$ such that $f - g$ is orthogonal to $W_{\mathcal{S}}$.
Finally, for any h in $W_{\mathcal{S}}$, $g - h$ is also in $W_{\mathcal{S}}$. Hence, $g - h$ is orthogonal to $f - g$. By the generalized Pythogoras theorem,

$$\|f - h\|^2 = \|(f - g) + (g - h)\|^2 = \|f - g\|^2 + \|g - h\|^2 \geq \|f - g\|^2.$$

This proves that

$$\|f - h\| \geq \|f - g\| \qquad \text{for all } h \in W_{\mathcal{S}}.$$

Now we restrict our consideration to functions f that are periodic of period 2π. In this case, the function is uniquely determined by its values on an interval $[a, b]$ of length 2π. We often take $I = [0, 2\pi]$ or $I = [-\pi, \pi]$. Notice that if $f : \mathbb{R} \to \mathbb{C}$ is a function of period 2π, then for any $\alpha \in \mathbb{R}$,

$$\int_{\alpha}^{\alpha+2\pi} f(x)dx = \int_{0}^{2\pi} f(x)dx = \int_{-\pi}^{\pi} f(x)dx.$$

Any function $f : [\alpha, \alpha + 2\pi] \to \mathbb{C}$ defined on an interval of length 2π can be extended to be a 2π-periodic function.

Definition 7.5 Extension of Functions

Let $I = [\alpha, \alpha + 2\pi]$ be an inverval of length 2π, and let $f : I \to \mathbb{C}$ be a function defined on I. We can extend f to be a 2π-periodic function $\widetilde{f} : \mathbb{R} \to \mathbb{C}$ in the following way.

(i) For $x \in (\alpha, \alpha + 2\pi)$, define

$$\widetilde{f}(x + 2n\pi) = f(x) \qquad \text{for all } n \in \mathbb{Z}.$$

(ii) For $x = \alpha$, define

$$\widetilde{f}(\alpha + 2n\pi) = \frac{f(\alpha) + f(\alpha + 2\pi)}{2} \qquad \text{for all } n \in \mathbb{Z}.$$

Examples are shown in Figures 7.1 and 7.2.

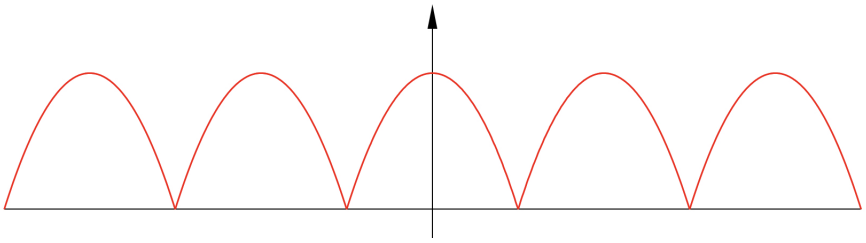


Figure 7.1: Extending a function $f : [-\pi, \pi] \to \mathbb{R}$ periodically.

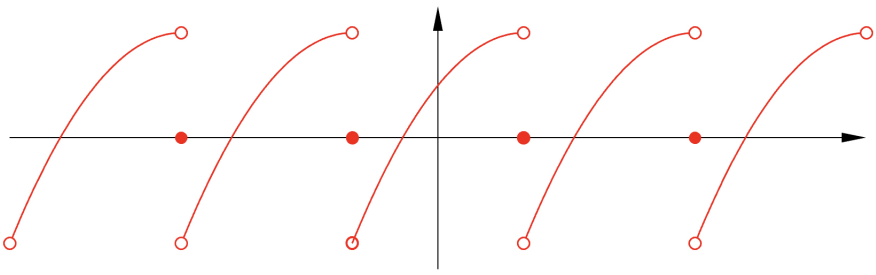


Figure 7.2: Extending a function $f : [-\pi, \pi] \to \mathbb{R}$ periodically.

Now let us define Fourier series. Example 7.2 asserts that the set

$$\mathcal{S} = \{\phi_n \,|\, n \in \mathbb{Z}\}, \quad \text{where } \phi_n(x) = e^{inx},$$

is an orthogonal system of functions in $\mathcal{R}(I,\mathbb{C})$, where $I=[-\pi,\pi]$. For $n\geq 0$, let W_n be the subspace of $\mathcal{R}(I,\mathbb{C})$ spanned by $\mathcal{S}_n=\left\{e^{ikx}\mid -n\leq k\leq n\right\}$. It is a vector space of dimension $2n+1$ with basis $\mathcal{S}_n$. Moreover,

$$W_0\subset W_1\subset W_2\subset\cdots.$$

A real basis of W_n is given by

$$\mathscr{B}_n=\{\sin kx\mid k=1,\ldots,n\}\cup\{\cos kx\mid k=0,1,\ldots,n\}.$$

Given $f\in\mathcal{R}(I,\mathbb{C})$, let $s_n=\text{proj}_{W_n}f$ be the projection of f onto W_n. The projection theorem says that

$$s_n(x)=\left(\text{proj}_{W_n}f\right)(x)=\sum_{k=-n}^{n}c_ke^{ikx},$$

where

$$c_k=\frac{\langle f,\phi_k\rangle}{\langle\phi_k,\phi_k\rangle}=\frac{1}{2\pi}\int_{-\pi}^{\pi}f(x)e^{-ikx}dx.$$

By Proposition 7.6 and the projection theorem, $s_n(x)$ can also be written as

$$s_n(x)=\left(\text{proj}_{W_n}f\right)(x)=\frac{a_0}{2}+\sum_{k=1}^{n}\left(a_k\cos kx+b_k\sin kx\right),$$

where

$$a_k=\frac{1}{\pi}\int_{-\pi}^{\pi}f(x)\cos kxdx,\qquad\text{for } 0\leq k\leq n,$$

$$b_k=\frac{1}{\pi}\int_{-\pi}^{\pi}f(x)\sin kxdx\qquad\text{for } 1\leq k\leq n.$$

By definition, we find that

$$c_0=\frac{a_0}{2},$$

and when $k\geq 1$,

$$c_{-k}=\frac{a_k+ib_k}{2},\quad c_k=\frac{a_k-ib_k}{2}.$$

If f is a real-valued function, a_k and b_k are real and $c_{-k}=\overline{c_k}$.

Definition 7.6 Trigonometric Series

A trigonometric series is a series of the form

$$\frac{a_0}{2}+\sum_{k=1}^{\infty}(a_k\cos kx+b_k\sin kx).$$

Since a trigonometric series can be expressed in the form $\sum_{k=-\infty}^{\infty} c_k e^{ikx}$, we also call a series of the form $\sum_{k=-\infty}^{\infty} c_k e^{ikx}$ a trigonometric series. Fourier series of a function is a trigonometric series.

Definition 7.7 Fourier Series and its n^{th} Partial Sums

Let $I = [-\pi, \pi]$. The Fourier series of a function f in $\mathcal{R}(I, \mathbb{C})$ is the infinite series

$$\sum_{k=-\infty}^{\infty} c_k e^{ikx} \quad \text{or} \quad \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx),$$

where

$$\begin{aligned} c_k &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-ikx} dx, && k \in \mathbb{Z}, \\ a_k &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos kx dx, && k \geq 0, \\ b_k &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin kx dx, && k \geq 1. \end{aligned}$$

The n^{th}-partial sum of the Fourier series is

$$s_n(x) = \sum_{k=-n}^{n} c_k e^{ikx} = \frac{a_0}{2} + \sum_{k=1}^{n} (a_k \cos kx + b_k \sin kx).$$

It is the projection of f onto the subspace of $\mathcal{R}(I, \mathbb{C})$ spanned by $\mathcal{S}_n = \{e^{ikx} \mid -n \leq k \leq n\}$.

Remark 7.1

If $f : \mathbb{R} \to \mathbb{C}$ is a 2π-periodic function which is Riemann integrable over a closed interval of length 2π, the Fourier series of f is the Fourier series of $f : [-\pi, \pi] \to \mathbb{C}$.

Remark 7.2

If $I = [-L, L]$, the Fourier series of a function $f \in \mathcal{R}(I, \mathbb{C})$ is the series

$$\sum_{k=-\infty}^{\infty} c_k \exp\left(\frac{i\pi kx}{L}\right),$$

where

$$c_k = \frac{1}{2L}\int_{-L}^{L} f(x)\exp\left(-\frac{i\pi kx}{L}\right)dx.$$

Henceforth, we only consider the case where I is a closed interval of length 2π. Let us look at some examples.

Example 7.4

Find the Fourier series of the function $f : [-\pi, \pi] \to \mathbb{R}$ defined as

$$f(x) = x.$$

Solution

When $k = 0$,

$$c_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)dx = \frac{1}{2\pi}\int_{-\pi}^{\pi} x dx = 0.$$

When $k \neq 0$,

$$\begin{aligned} c_k &= \frac{1}{2\pi}\int_{-\pi}^{\pi} xe^{-ikx}dx \\ &= \frac{1}{2\pi}\left\{\left[-\frac{1}{ik}xe^{-ikx}\right]_{-\pi}^{\pi} + \frac{1}{ik}\int_{-\pi}^{\pi} e^{-ikx}dx\right\} \\ &= \frac{(-1)^{k-1}}{ik}. \end{aligned}$$

Therefore, the Fourier series of f is

$$\sum_{k=1}^{\infty}(-1)^{k-1}\frac{e^{ikx} - e^{-ikx}}{ik} = \sum_{k=1}^{\infty} 2(-1)^{k-1}\frac{\sin kx}{k}.$$

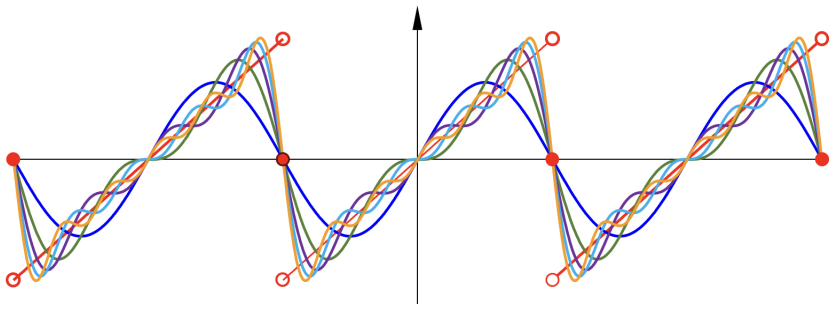


Figure 7.3: The function $f(x) = x, -\pi < x < \pi$ and $s_n(x)$, $1 \leq n \leq 5$.

Remark 7.3

Let $I = [-\pi, \pi]$. Given f in $\mathcal{R}(I, \mathbb{C})$, we call each

$$c_k = \int_I f(x)e^{-ikx}dx, \qquad k \in \mathbb{Z}$$

a Fourier coefficient of f. The mapping $\mathfrak{F}_k$ from $\mathcal{R}(I, \mathbb{C})$ to $\mathbb{C}$ which takes a function f to c_k is a linear transformation between vector spaces.
When $f : I \to \mathbb{R}$ is a real-valued function, we usually prefer to work with the Fourier coefficients a_k and b_k. One can show that if $f : [-\pi, \pi] \to \mathbb{R}$ is an odd function, then $a_k = 0$ for all $k \geq 0$, so that the Fourier series of f only has sine terms. If $f : [-\pi, \pi] \to \mathbb{R}$ is an even function, then $b_k = 0$ for all $k \geq 1$, so that the Fourier series of f only has the constant and the cosine terms.

Remark 7.4

If $f : I \to \mathbb{C}$ is a function of the form

$$f(x) = \sum_{k \in J} c_k e^{ikx},$$

where J is a finite subset of integers, then the Fourier series of f is equal to itself.

Example 7.5

The Fourier series of a constant function $f : I \to \mathbb{C}$, $f(x) = c$ is just c itself.

Example 7.6

Let $f : [0, 2\pi] \to \mathbb{R}$ be the function defined as

$$f(x) = x(2\pi - x).$$

Find its Fourier series, and express it in terms of trigonometric functions.

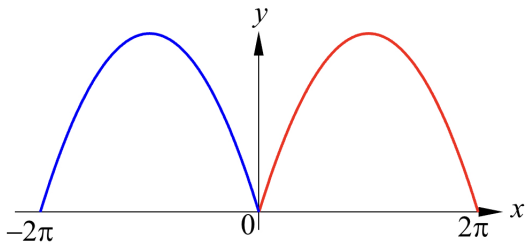


Figure 7.4: The function $f : [0, 2\pi] \to \mathbb{R}$, $f(x) = x(2\pi - x)$ and its entension.

Solution

When we extend f periodically to the function $\widetilde{f} : \mathbb{R} \to \mathbb{R}$, we find that when $x \in [0, 2\pi]$,

$$\widetilde{f}(-x) = f(-x + 2\pi) = (2\pi - x)x = f(x) = \widetilde{f}(x).$$

Hence, $\widetilde{f}(x)$ is an even function. This implies that the Fourier series of $f : [0, 2\pi] \to \mathbb{R}$ only has cosine terms.

$$a_0 = \frac{1}{\pi}\int_0^{2\pi} f(x)dx = \frac{2}{\pi}\int_0^{\pi}(2\pi x - x^2)dx = \frac{2}{\pi}\left[\pi x^2 - \frac{x^3}{3}\right]_0^{\pi} = \frac{4}{3}\pi^2.$$

For $k \geq 1$,

$$\begin{aligned} a_k &= \frac{2}{\pi}\int_0^\pi f(x)\cos x dx = \frac{2}{\pi}\int_0^\pi (2\pi x - x^2)\cos x dx \\ &= \frac{2}{\pi}\left(\left[\frac{(2\pi x - x^2)\sin kx}{k}\right]_0^\pi - \frac{1}{k}\int_0^\pi (2\pi - 2x)\sin kx dx\right) \\ &= -\frac{2}{\pi k}\left(\left[-\frac{(2\pi - 2x)\cos kx}{k}\right]_0^\pi - \frac{2}{k}\int_0^\pi \cos kx dx\right) \\ &= -\frac{4}{k^2} + \frac{4}{\pi k^3}\left[\sin kx\right]_0^\pi = -\frac{4}{k^2}. \end{aligned}$$

Therefore, the Fourier series of f is

$$\frac{2}{3}\pi^2 - 4\sum_{k=1}^{\infty}\frac{\cos kx}{k^2}.$$

Example 7.7

Let a and b be two numbers satisfying $-\pi \leq a < b \leq \pi$, and let $g : [-\pi, \pi] \to \mathbb{R}$ be the function defined as

$$g(x) = \begin{cases} 1, & \text{if } a \leq x \leq b, \\ 0, & \text{otherwise.} \end{cases}$$

Find the Fourier series of g in exponential form.

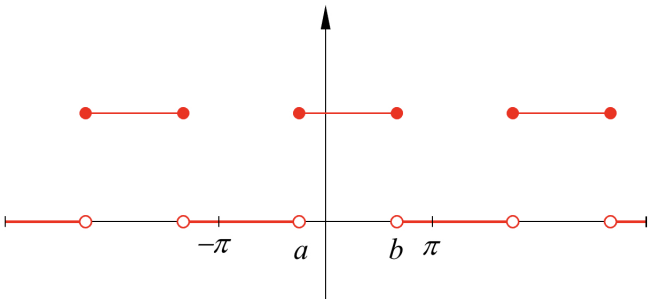


Figure 7.5: The function $g : [0, 2\pi] \to \mathbb{R}$ defined in Example 7.7 and its entension.

Solution

Since g is piecewise continuous, it is Riemann integrable.

$$c_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} g(x)dx = \frac{1}{2\pi}\int_a^b dx = \frac{b-a}{2\pi}.$$

For $k \geq 1$,

$$c_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} g(x)e^{-ikx}dx = \frac{1}{2\pi}\int_a^b e^{-ikx}dx = \frac{e^{-ikb}-e^{-ika}}{-2\pi ik},$$

$$c_{-k} = \overline{c_k} = \frac{e^{ikb}-e^{ika}}{2\pi ik}.$$

Therefore, the Fourier series of g is

$$\frac{b-a}{2\pi} + \frac{i}{2\pi}\sum_{k=1}^{\infty}\frac{(e^{-ikb}-e^{-ika})e^{ikx}-(e^{ikb}-e^{ika})e^{-ikx}}{k}.$$

Example 7.8

Find the Fourier series of the function $f : [-\pi, \pi] \to \mathbb{R}$, $f(x) = x\sin x$.

Solution

Since f is a real-valued even function, we only need to compute the Fourier coefficients

$$a_k(f) = \frac{1}{\pi}\int_{-\pi}^{\pi} x\sin x\cos kx dx \qquad \text{when } k \geq 0.$$

Let $g : [-\pi, \pi] \to \mathbb{R}$ be the function $g(x) = x$. We have seen in Example 7.4 that the Fourier series of g is given by

$$G(x) = \sum_{k=1}^{\infty} 2(-1)^{k-1}\frac{\sin kx}{k}.$$

Therefore, for $k \geq 1$,

$$b_k(g) = \frac{1}{\pi}\int_{-\pi}^{\pi} x\sin kx = \frac{2(-1)^{k-1}}{k}.$$

From this, we find that

$$a_0(f) = b_1(g) = 2,$$
$$a_1(f) = \frac{1}{2\pi}\int_{-\pi}^{\pi} x\sin 2x dx = \frac{1}{2}b_2(g) = -\frac{1}{2};$$

and when $k \geq 2$,

$$\begin{aligned} a_k(f) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} x\left(\sin(k+1)x - \sin(k-1)x\right)dx \\ &= \frac{1}{2}\left(b_{k+1}(g) - b_{k-1}(g)\right) \\ &= (-1)^k\left(\frac{1}{k+1} - \frac{1}{k-1}\right) \\ &= \frac{2(-1)^{k-1}}{k^2-1}. \end{aligned}$$

Hence, the Fourier series of the function $f : [-\pi, \pi] \to \mathbb{R}$, $f(x) = x\sin x$ is

$$1 - \frac{1}{2}\cos x + \sum_{k=2}^{\infty}\frac{2(-1)^{k-1}}{k^2-1}\cos kx.$$

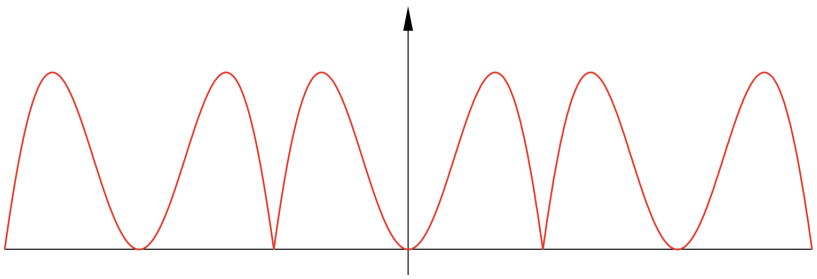


Figure 7.6: The function $f : [-\pi, \pi] \to \mathbb{R}$, $f(x) = x\sin x$ and its periodic entension.

At the end of this section, let us make an additional remark.

Remark 7.5 Semi-Norms

A semi-norm on a complex vector space V is a function $\| \cdot \| : V \to \mathbb{R}$ which defines the norm $\|\mathbf{v}\|$ for each $\mathbf{v}$ in V such that the following hold.

(a) For any $\mathbf{v} \in V$, $\|\mathbf{v}\| \geq 0$.

(b) For any $\alpha \in \mathbb{C}$, and any $\mathbf{v} \in V$, $\|\alpha\mathbf{v}\| = |\alpha|\|\mathbf{v}\|$.

(c) For any $\mathbf{u}$ and $\mathbf{v}$ in V, $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$.

If in addition, we have

(d) $\|\mathbf{v}\| = 0$ if and only if $\mathbf{v} = 0$,

then $\| \cdot \|$ is called a norm on the vector space V.
Proposition 7.2 and Proposition 7.4 justify that the L^2-norm

$$\|f\|_2 = \sqrt{\int_I |f(x)|^2 dx}$$

is indeed a semi-norm on the vector space $\mathcal{R}(I, \mathbb{C})$.
There are other semi-norms on $\mathcal{R}(I, \mathbb{C})$. One of them which will also be useful later is the L^1-norm defined as

$$\|f\|_1 = \int_I |f(x)| dx.$$

The fact that this is a semi-norm is quite easy to establish.

Exercises 7.1

Question 1

Let $f : [-\pi, \pi] \to \mathbb{R}$ be a real-valued Riemann integrable function.

(a) If $f : [-\pi, \pi] \to \mathbb{R}$ is an odd function, show that the Fourier series of f has the form

$$\sum_{k=1}^{\infty} b_k \sin kx, \qquad \text{where } b_k = \frac{2}{\pi} \int_0^{\pi} f(x) \sin kx.$$

(b) If $f : [-\pi, \pi] \to \mathbb{R}$ is an even function, show that the Fourier series of f has the form

$$\frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos kx, \qquad \text{where } a_k = \frac{2}{\pi} \int_0^{\pi} f(x) \cos kx.$$

Question 2

Find the Fourier series of the function $f : [-\pi, \pi] \to \mathbb{R}$, $f(x) = |x|$, and express it in terms of trigonometric functions.

Question 3

Find the Fourier series of the function $f : [-\pi, \pi] \to \mathbb{R}$, $f(x) = x^2$, and express it in terms of trigonometric functions.

Question 4

Find the Fourier series of the function $f : [0, 2\pi] \to \mathbb{R}$, $f(x) = x^2$, and express it in terms of trigonometric functions.

Question 5

Find the Fourier series of the function $f : [-\pi, \pi] \to \mathbb{R}$, $f(x) = \sin 2x$, and express it in terms of trigonometric functions.

Question 6

Find the Fourier series of the function $f : [-\pi, \pi] \to \mathbb{R}$,

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

and express it in terms of trigonometric functions.

Question 7

Find the Fourier series of the function $f : [-\pi, \pi] \to \mathbb{R}$, $f(x) = x \cos x$ from the Fourier series of the function $g : [-\pi, \pi] \to \mathbb{R}$, $g(x) = x$.

Question 8

Let x_0 be a point in the interval $[a, b]$, and let $f : [a, b] \to \mathbb{C}$ and $g : [a, b] \to \mathbb{C}$ be L^2-equivalent Riemann integrable functions. Assume that both f and g are continuous at the point x_0, show that $f(x_0) = g(x_0)$.

7.2 The Pointwise Convergence of a Fourier Series

Let $I = [-\pi, \pi]$ and let $\mathcal{R}(I, \mathbb{C})$ be the vector space that consists of all Riemann integrable functions $f : I \to \mathbb{C}$ that are defined on I. Each of these functions can be extended to a periodic function $\widetilde{f} : \mathbb{R} \to \mathbb{C}$ so that $\widetilde{f}(x) = f(x)$ for all x in the interior of I.

Given $f \in \mathcal{R}(I, \mathbb{C})$, we define the Fourier series of f as the infinite series

$$\sum_{k=-\infty}^{\infty} c_k e^{ikx} = \frac{a_0}{2} + \sum_{k=1}^{\infty} \left(a_k \cos kx + b_k \sin kx \right),$$

where

$$\begin{aligned} c_k &= \frac{1}{2\pi} \int_I f(x) e^{-ikx} dx, \qquad k \in \mathbb{Z}, \\ a_k &= \frac{1}{\pi} \int_I f(x) \cos kx dx, \qquad k \geq 0, \\ b_k &= \frac{1}{\pi} \int_I f(x) \sin kx dx, \qquad k \geq 1. \end{aligned}$$

The problem of interest to us is the convergence of the Fourier series. Given $n \geq 0$, the n^{th}-partial sum of the Fourier series of f is

$$s_n(x) = \sum_{k=-n}^{n} c_k e^{ikx}.$$

We say that the Fourier series converges pointwise if the sequence of partial sum functions $\{s_n : I \to \mathbb{C}\}$ converges pointwise. Let us first give an integral expression for the partial sums $s_n(x)$ of the Fourier series. By definition,

$$\begin{aligned} s_n(x) &= \frac{1}{2\pi} \sum_{k=-n}^{n} \left(\int_{-\pi}^{\pi} f(t) e^{-ikt} dt \right) e^{ikx} \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \sum_{k=-n}^{n} e^{ik(x-t)} dt. \end{aligned}$$

If $x \in 2\pi\mathbb{Z}$, $e^{ikx} = 1$ for all $-n \leq k \leq n$. Therefore,

$$\sum_{k=-n}^{n} e^{ikx} = 2n + 1.$$

If $x \notin 2\pi\mathbb{Z}$, $e^{ix} \neq 1$. Using the sum formula for a geometric sequence, we have

$$\sum_{k=-n}^{n} e^{ikx} = e^{-inx}\left(1 + e^{ix} + \cdots + e^{2inx}\right) = \frac{e^{-i\left(n+\frac{1}{2}\right)x}}{e^{-\frac{ix}{2}}} \times \frac{e^{i(2n+1)x} - 1}{e^{ix} - 1}$$

$$= \frac{e^{i(n+\frac{1}{2})x} - e^{-i(n+\frac{1}{2})x}}{e^{\frac{ix}{2}} - e^{-\frac{ix}{2}}} = \frac{\sin\left(n + \frac{1}{2}\right)x}{\sin\frac{x}{2}}.$$

Definition 7.8 Dirichlet Kernel

Given a nonnegative integer n, the Dirichlet kernel $D_n : \mathbb{R} \to \mathbb{R}$ is

$$D_n(x) = \sum_{k=-n}^{n} e^{ikx} = \begin{cases} 2n+1, & \text{if } x \in 2\pi\mathbb{Z}, \\ \dfrac{\sin\left(n + \frac{1}{2}\right)x}{\sin\frac{x}{2}}, & \text{otherwise.} \end{cases}$$

Our derivation above gives the following.

Proposition 7.8

Let $I = [-\pi, \pi]$, and let $f : I \to \mathbb{C}$ be a Riemann integrable function. The n^{th}-partial sum $s_n(x)$ of the Fourier series of f has an integral representation given by

$$s_n(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)D_n(x-t)dt,$$

where $D_n : \mathbb{R} \to \mathbb{R}$ is the Dirichlet kernel given by

$$D_n(x) = \begin{cases} 2n+1, & \text{if } x \in 2\pi\mathbb{Z}, \\ \dfrac{\sin\left(n + \frac{1}{2}\right)x}{\sin\frac{x}{2}}, & \text{otherwise.} \end{cases}$$

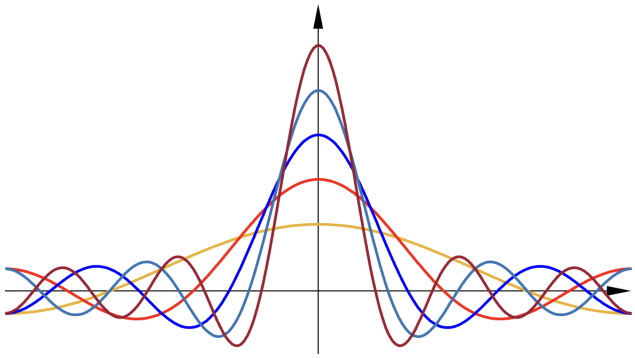


Figure 7.7: The Dirichlet kernels $D_n : [-\pi, \pi] \to \mathbb{R}$ for $1 \leq n \leq 5$.

Remark 7.6

By definition, the Dirichlet kernel $D_n(x)$ is equal to

$$D_n(x) = \sum_{k=-n}^{n} e^{ikx}.$$

From this, one can see that $D_n(x)$ is an infinitely differentiable 2π-periodic function, and it is an even function.

Recall that $g : [a, b] \to \mathbb{C}$ is a step function if there is a partition $P = \{x_0, x_1, \ldots, x_l\}$ of $[a, b]$ such that for each $1 \leq j \leq l$, $g : (x_{j-1}, x_j) \to \mathbb{C}$ is a constant function. It is easy to see that $g : [a, b] \to \mathbb{C}$ is a step function if and only if both its real and imaginary parts are step functions. The following theorem asserts that a Riemann integrable function $f : [a, b] \to \mathbb{R}$ can be approximated in L^1 by step functions.

Theorem 7.9

Let $f : [a, b] \to \mathbb{C}$ be a Riemann integrable function. For every $\varepsilon > 0$, there is a step function $g : [a, b] \to \mathbb{C}$ such that

$$\int_a^b |f(x) - g(x)|dx < \varepsilon.$$

Proof

Let $f(x) = u(x) + iv(x)$, where $u : [a,b] \to \mathbb{R}$ and $v : [a,b] \to \mathbb{R}$ are the real and imaginary parts of f. Assume that $u_1 : [a,b] \to \mathbb{R}$ and $v_1 : [a,b] \to \mathbb{R}$ are step functions such that

$$\int_a^b |u(x) - u_1(x)|dx < \frac{\varepsilon}{2}, \qquad \int_a^b |v(x) - v_1(x)|dx < \frac{\varepsilon}{2}.$$

Let $g : [a,b] \to \mathbb{R}$ be the function $g = u_1 + iv_1$. Then g is a step function. By triangle inequality, we have

$$\int_a^b |f(x) - g(x)|dx \leq \int_a^b |u(x) - u_1(x)|dx + \int_a^b |v(x) - v_1(x)|dx < \varepsilon.$$

Therefore, it is sufficient to prove the theorem when f is a real-valued function.

Given $\varepsilon > 0$, since $f : [a,b] \to \mathbb{R}$ is Riemann integrable, there is a partition $P = \{x_0, x_1, \ldots, x_l\}$ of $[a,b]$ such that

$$U(f,P) - L(f,P) < \varepsilon,$$

where

$$U(f,P) = \sum_{j=1}^{l} M_j(x_j - x_{j-1}), \qquad M_j = \sup_{x_{j-1} \leq x \leq x_j} f(x),$$

and

$$L(f,P) = \sum_{j=1}^{l} m_j(x_j - x_{j-1}), \qquad m_j = \inf_{x_{j-1} \leq x \leq x_j} f(x),$$

are respectively the Darboux upper sum and Darboux lower sum of f with respect to the partition P. Define the function $g : [a,b] \to \mathbb{R}$ by

$$g(x) = m_j \qquad \text{when } x_{j-1} \leq x < x_j,$$

and $g(b) = f(b)$. Then g is a step function, and

$$f(x) \geq g(x) \qquad \text{for all } x \in [a,b].$$

It follows that

$$\begin{aligned}\int_a^b |f(x)-g(x)|dx &= \int_a^b (f(x)-g(x))dx \\ &= \int_a^b f(x)dx - \sum_{j=1}^{l}\int_{x_{j-1}}^{x_j} g(x)dx \\ &\leq U(f,P) - \sum_{j=1}^{l} m_j(x_j - x_{j-1}) \\ &= U(f,P) - L(f,P) < \varepsilon.\end{aligned}$$

This completes the proof.

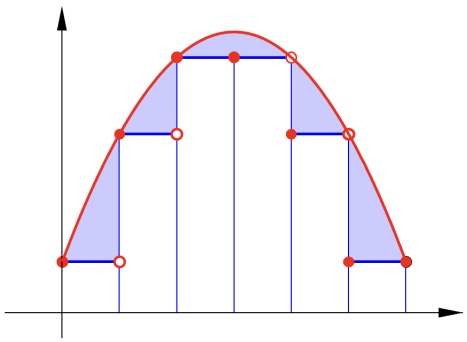


Figure 7.8: Approximating a Riemann integrable function by a step function.

An important tool in the proof of pointwise convergence of Fourier series is the Riemann-Lebesgue lemma. This lemma is important in its own right. Hence, we state it in the most general setting.

Theorem 7.10 The Riemann-Lebesgue Lemma

Let $I = [a, b]$. If $f : I \to \mathbb{C}$ is a Riemann integrable function, then

$$\lim_{\beta\to\infty}\int_a^b f(x)e^{i\beta x}dx = 0.$$

Proof

Given $\varepsilon > 0$, Theorem 7.9 says that there is a step function $g : [a, b] \to \mathbb{R}$ such that

$$\int_a^b |f(x) - g(x)|dx < \frac{\varepsilon}{2}.$$

This implies that

$$\left|\int_a^b (f(x) - g(x))e^{i\beta x}dx\right| \leq \int_a^b |f(x) - g(x)|dx < \frac{\varepsilon}{2}.$$

Let $P = \{x_0, x_1, \ldots, x_l\}$ be a partition of $[a, b]$ such that for $1 \leq j \leq l$, $g(x) = m_j$ for all x in (x_{j-1}, x_j). Let $M = \max\{|m_1|, \ldots, |m_l|\}$. Then

$$\left|\int_{x_{j-1}}^{x_j} g(x)e^{i\beta x}dx\right| = \left|\frac{m_j}{i\beta}\left(e^{i\beta x_j} - e^{i\beta x_{j-1}}\right)\right| \leq \frac{2M}{\beta}.$$

It follows that if $\beta > \dfrac{4Ml}{\varepsilon}$,

$$\left|\int_a^b g(x)e^{i\beta x}dx\right| \leq \sum_{j=1}^{l}\left|\int_{x_{j-1}}^{x_j} g(x)e^{i\beta x}dx\right| \leq \frac{2Ml}{\beta} < \frac{\varepsilon}{2}.$$

Therefore,

$$\left|\int_a^b f(x)e^{i\beta x}dx\right| \leq \left|\int_a^b (f(x) - g(x))e^{i\beta x}dx\right| + \left|\int_a^b g(x)e^{i\beta x}dx\right| < \varepsilon.$$

This proves the assertion.

Since

$$\sin \beta x = \frac{e^{i\beta x} - e^{-i\beta x}}{2i},$$

we obtain the following.

Corollary 7.11

Let $I = [a, b]$. If $f : I \to \mathbb{C}$ is a Riemann integrable function, then

$$\lim_{\beta\to\infty} \int_a^b f(x) \sin \beta x dx = 0.$$

Recall that a function $f : [a, b] \to \mathbb{C}$ is piecewise continuous if there is a partition $P = \{x_0, x_1, \ldots, x_l\}$ of $[a, b]$ such that for each $1 \leq j \leq l$, $f : (x_{j-1}, x_j) \to \mathbb{C}$ is a continuous functions. It is piecewise differentiable if there is a partition $P = \{x_0, x_1, \ldots, x_l\}$ of $[a, b]$ such that for each $1 \leq j \leq l$, $f : (x_{j-1}, x_j) \to \mathbb{C}$ is a differentiable functions. Obviously, if $f : [a, b] \to \mathbb{C}$ is piecewise differentiable, it is piecewise continuous. The piecewise continuity and piecewise differentiability do not impose any conditions on the partition points. In the following, we introduce a class of functions which satisfy stronger conditions on the partition points.

Definition 7.9 Strongly Piecewise Differentiable Functions

A function $f : [a, b] \to \mathbb{C}$ is strongly piecewise continuous if there is a partition $P = \{x_0, x_1, \ldots, x_l\}$ of $[a, b]$ such that for each $1 \leq j \leq l$, the limits

$$f_+(x_{j-1}) = \lim_{x \to x_{j-1}^+} f(x) \quad \text{and} \quad f_-(x_j) = \lim_{x \to x_j^-} f(x)$$

exist, and the function $g_j : [x_{j-1}, x_j] \to \mathbb{C}$ defined as

$$g_j(x) = \begin{cases} f_+(x_{j-1}), & \text{if } x = x_{j-1} \\ f(x), & \text{if } x_{j-1} < x < x_j \\ f_-(x_j), & \text{if } x = x_j \end{cases}$$

is continuous. If f is also differentiable on (x_{j-1}, x_j), and the limits

$$f'_+(x_{j-1}) = \lim_{h \to 0^+} \frac{f(x_{j-1} + h) - f_+(x_{j-1})}{h}$$

and

$$f'_-(x_j) = \lim_{h \to 0=^+} \frac{f_-(x_j) - f(x_j - h)}{h}$$

exist, we say that $f : [a, b] \to \mathbb{C}$ is strongly piecewise differentiable.

Notice that a strongly piecewise differentiable function is strongly piecewise continuous and bounded. Therefore, it is Riemann integrable.

We have abused notation above and denote the limit

$$\lim_{h \to 0^+} \frac{f(c + h) - f_+(c)}{h}$$

as $f'_+(c)$. Strictly speaking, $f'_+(c)$ is the right derivative of f at c which is defined as

$$\lim_{h\to 0^+} \frac{f(c+h)-f(c)}{h}.$$

The two expressions are equivalent if $f(c) = f_+(c)$, meaning that f is right continuous at c. In fact, for the limit

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

to exist, a necessary condition is $f_+(c)$ exists and is equal to $f(c)$. However, here we do not require the function f to be continuous at the partition points x_j, $0 \le j \le l$. We only require the function to have left and right limits at these points. Since $f(c) = f_+(c) = f_-(c)$ only when f is continuous at c, we modify the definitions of $f'_+(c)$ and $f'_-(c)$ for functions that can have discontinuity at the point c.

If x is an interior point of I and $f : I \to \mathbb{C}$ is differentiable at x,

$$\lim_{h\to 0^+} \frac{f(x+h)-f(x)}{h} = f'_+(x) = f'(x) = f'_-(x) = \lim_{h\to 0^+} \frac{f(x)-f(x-h)}{h}.$$

Thus, if the function $f : [a,b] \to \mathbb{C}$ is strongly piecewise differentiable, then for any $x \in (a,b)$,

$$\lim_{h\to 0^+} \frac{f(x+h)+f(x-h)-f_+(x)-f_-(x)}{h} = f'_+(x) - f'_-(x). \tag{7.2}$$

Example 7.9

The function $f : [-\pi,\pi] \to \mathbb{R}$ defined as

$$f(x) = \begin{cases} \pi - x, & \text{if } -\pi \le x < 0, \\ x^2, & \text{if } 0 \le x \le \pi, \end{cases}$$

is strongly piecewise differentiable.

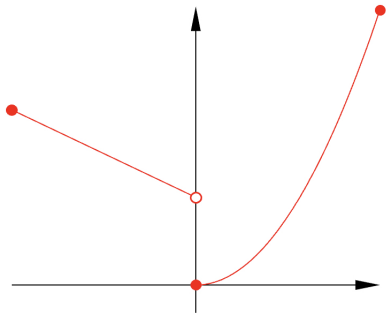


Figure 7.9: The strongly piecewise differentiable function defined in Example 7.9.

Lemma 7.12

Let $f : [-\pi, \pi] \to \mathbb{C}$ be a strongly piecewise differentiable function, and let $\widetilde{f} : \mathbb{R} \to \mathbb{C}$ be the 2π-periodic extension of f. Given $x \in \mathbb{R}$, define the function $h : [0, \pi] \to \mathbb{C}$ by

$$h(t) = \frac{\widetilde{f}(x+t) + \widetilde{f}(x-t) - \widetilde{f}_+(x) - \widetilde{f}_-(x)}{\sin \dfrac{t}{2}}, \quad t \in (0, \pi], \tag{7.3}$$

and $h(0)$ can be any value. Then $h : [0, \pi] \to \mathbb{C}$ is a piecewise continuous bounded function. Hence, $h : [0, \pi] \to \mathbb{C}$ is Riemann integrable.

Proof

When $t \in [0, \pi]$, $\sin \dfrac{t}{2} = 0$ only when $t = 0$. Notice that $\widetilde{f}$ is a bounded piecewise continuous function. Hence, $h : [0, \pi] \to \mathbb{C}$ is piecewise continuous. For any positive number r that is less than π, h is bounded on $[r, \pi]$. To show that $h : [0, \pi] \to \mathbb{C}$ is bounded, it is sufficient to show that $h(t)$ has a limit when $t \to 0^+$. Now

$$\lim_{t \to 0^+} h(t) = \lim_{t \to 0^+} \frac{\widetilde{f}(x+t) + \widetilde{f}(x-t) - \widetilde{f}_+(x) - \widetilde{f}_-(x)}{t} \lim_{t \to 0^+} \frac{t}{\sin \dfrac{t}{2}}.$$

Since

$$\lim_{t\to 0^+} \frac{t}{\sin\frac{t}{2}} = 2,$$

eq. (7.2) gives

$$\lim_{t\to 0^+} h(t) = 2\left(\widetilde{f}'_+(x) - \widetilde{f}'_-(x)\right).$$

This completes the proof.

Now we can prove the Dirichlet's theorem.

Theorem 7.13 Dirichlet's Theorem

Let $f : [-\pi, \pi] \to \mathbb{C}$ be a strongly piecewise differentiable function, and let $\widetilde{f} : \mathbb{R} \to \mathbb{C}$ be the 2π-periodic extension of f. For every $x \in \mathbb{R}$, the Fourier series of f converges at the point x to

$$\frac{\widetilde{f}_-(x) + \widetilde{f}_+(x)}{2}.$$

Proof

Notice that the function $\widetilde{f}$ is also strongly piecewise differentiable on any compact interval. The n^{th}-partial sum of the Fourier series of f is

$$s_n(x) = \sum_{k=-n}^{n} c_k e^{ikx}, \qquad \text{where } c_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-ikx}dx.$$

For a fixed real number x, we want to show that $s_n(x)$ converges to the number

$$u = \frac{\widetilde{f}_-(x) + \widetilde{f}_+(x)}{2}.$$

By Proposition 7.8,

$$s_n(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)D_n(x-t)dt = \frac{1}{2\pi}\int_{x-\pi}^{x+\pi} \widetilde{f}(x-t)D_n(t)dt.$$

Since $\widetilde{f}$ and D_n are 2π-periodic functions, we find that

$$s_n(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \widetilde{f}(x-t)D_n(t)dt.$$

Notice that

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} D_n(t)dt = \frac{1}{2\pi}\sum_{k=-n}^{n}\int_{-\pi}^{\pi} e^{ikt}dt = 1.$$

Therefore,

$$s_n(x) - u = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\widetilde{f}(x-t) - u\right)D_n(t)dt.$$

Using the fact that $D_n : \mathbb{R} \to \mathbb{R}$ is an even function, we find that

$$\begin{aligned} s_n(x) - u &= \frac{1}{2\pi}\int_{0}^{\pi}\left(\widetilde{f}(x+t) + \widetilde{f}(x-t) - 2u\right)D_n(t)dt \\ &= \frac{1}{2\pi}\int_{0}^{\pi}\left(\widetilde{f}(x+t) + \widetilde{f}(x-t) - 2u\right)\frac{\sin\left(n+\frac{1}{2}\right)t}{\sin\frac{t}{2}}dt \\ &= \frac{1}{2\pi}\int_{0}^{\pi} h(t)\sin\left[\left(n+\frac{1}{2}\right)t\right]dt, \end{aligned}$$

where $h : [0,\pi] \to \mathbb{C}$ is the function defined by (7.3). By Lemma 7.12, $h : [0,\pi] \to \mathbb{C}$ is Riemann integrable. By the Riemann-Lebesgue lemma,

$$\lim_{n\to\infty}\int_{0}^{\pi} h(t)\sin\left[\left(n+\frac{1}{2}\right)t\right]dt = 0.$$

This proves that

$$\lim_{n\to\infty} s_n(x) = u.$$

From the Dirichlet's theorem, we can spell out the following explicitly.

Corollary 7.14

If $f : [-\pi, \pi] \to \mathbb{C}$ is a strongly piecewise differentiable function, then its Fourier series converges pointwise. Denote by $F : \mathbb{R} \to \mathbb{C}$ the Fourier series of f. Then for any $x \in \mathbb{R}$,

$$F(x) = \frac{\widetilde{f}_+(x) + \widetilde{f}_-(x)}{2},$$

where $\widetilde{f} : \mathbb{R} \to \mathbb{C}$ is the 2π-periodic extension of f. This implies that

(a) If $x \in (-\pi, \pi)$ and f is continuous at x, then $F(x) = f(x)$.

(b) If f is right continuous at $-\pi$, left continuous at π, and $f(-\pi) = f(\pi)$, then $F(-\pi) = F(\pi) = f(-\pi) = f(\pi)$.

Let us look at a few examples.

Example 7.10

Let $f : [-\pi, \pi] \to \mathbb{R}$ be the function $f(x) = x$ considered in Example 7.4. Since f is a strongly differentiable function, the Fourier series of f converges pointwise. We have shown that the Fourier series is given by

$$F(x) = \sum_{k=1}^{\infty} 2(-1)^{k-1} \frac{\sin kx}{k}.$$

For any $x \in (-\pi, \pi)$, this series converges to x. When $x = \pi$, it converges to

$$0 = \frac{f_+(-\pi) + f_-(\pi)}{2}.$$

When $x = \dfrac{\pi}{2}$, since $\sin \dfrac{\pi k}{2}$ is 0 when $k = 2n$, it is equal to 1 when $k = 4n + 1$, and it is equal to -1 when $k = 4n + 3$, we deduce that

$$\frac{1}{2} F\left(\frac{\pi}{2}\right) = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots = \frac{1}{2} f\left(\frac{\pi}{2}\right) = \frac{\pi}{4},$$

which is just the Newton-Gregory formula.

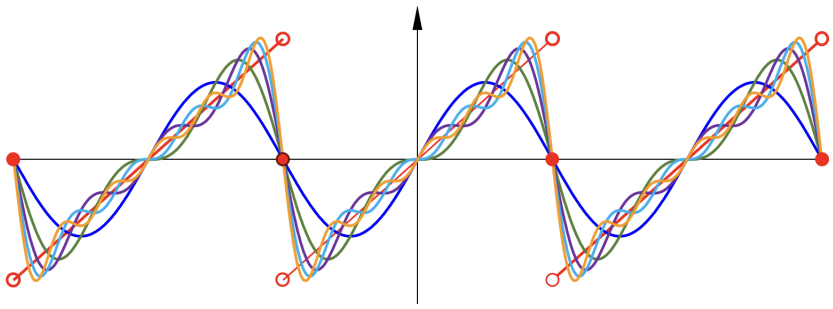


Figure 7.10: Convergence of the Fourier series of the function $f(x) = x, -\pi < x < \pi$.

Example 7.11

The function $f : [0, 2\pi] \to \mathbb{R}$, $f(x) = x(2\pi - x)$ considered in Example 7.5 is a strongly piecewise differentiable function. Hence, its Fourier series

$$F(x) = \frac{2}{3}\pi^2 - 4\sum_{k=1}^{\infty} \frac{\cos kx}{k^2}$$

converges everywhere. Since $f : [-\pi, \pi] \to \mathbb{C}$ is continuous and $f(-\pi) = f(\pi)$, we find that $F(x) = f(x)$ for all $x \in [-\pi, \pi]$. In particular, setting $x = 0$ and $x = \pi$ respectively, we find that

$$F(0) = \frac{2\pi^2}{3} - 4\sum_{k=1}^{\infty} \frac{1}{k^2} = f(0) = 0,$$

$$F(\pi) = \frac{2\pi^2}{3} - 4\sum_{k=1}^{\infty} \frac{(-1)^k}{k^2} = f(\pi) = \pi^2.$$

These give

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \frac{1}{6^2} + \cdots = \frac{\pi^2}{6},$$

$$\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k^2} = 1 - \frac{1}{2^2} + \frac{1}{3^2} - \frac{1}{4^2} + \frac{1}{5^2} - \frac{1}{6^2} + \cdots = \frac{\pi^2}{12}.$$

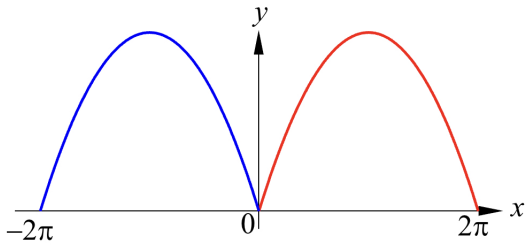


Figure 7.11: The function $f : [0, 2\pi] \to \mathbb{R}$, $f(x) = x(2\pi - x)$ and its entension.

Example 7.12

In Example 7.7, we consider the function $g : [-\pi, \pi] \to \mathbb{R}$ defined as

$$g(x) = \begin{cases} 1, & \text{if } a \le x \le b, \\ 0, & \text{otherwise,} \end{cases}$$

where a and b are two numbers satisfying $-\pi \le a < b \le \pi$. Notice that g is a strongly piecewise differentiable function. Thus, its Fourier series

$$G(x) = \frac{b-a}{2\pi} + \frac{i}{2\pi} \sum_{k=1}^{\infty} \frac{(e^{-ikb} - e^{-ika})e^{ikx} - (e^{ikb} - e^{ika})e^{-ikx}}{k}$$

converges pointwise. If $x \in (a, b)$, Dirichlet's theorem says that $G(x) = g(x) = 1$. Hence, for any $x \in (a, b) \subset (-\pi, \pi)$,

$$b - a = 2\pi - i \sum_{k=1}^{\infty} \sum_{k=1}^{\infty} \frac{(e^{-ikb} - e^{-ika})e^{ikx} - (e^{ikb} - e^{ika})e^{-ikx}}{k}.$$

Remark 7.7

If we scrutinize the proof of Theeorem 7.13, we find that a necessary and sufficient condition for the Fourier series of a 2π-periodic function $f : \mathbb{R} \to \mathbb{C}$ to converge at a point x is that the limit

$$\lim_{n\to\infty} \int_0^\pi \frac{f(x+t)+f(x-t)}{t} \sin\left[\left(n+\frac{1}{2}\right)t\right] dt$$

should exists. This is known as the *Riemann's localization theorem*. Theeorem 7.13 says that if $f : [-\pi, \pi] \to \mathbb{R}$ is strongly piecewise differentiable, then this limit exists. This is sufficient for most of our applications.

Remark 7.8 Fourier Sine Series and Fourier Cosine Series

Let $L > 0$, and let $f : [0, L] \to \mathbb{C}$ be a Riemann integrable function defined on $[0, L]$. We can extend f to be an odd function $f_o : [-L, L] \to \mathbb{C}$ by defining

$$f_o(x) = \begin{cases} -f(-x), & \text{if } -L \le x < 0, \\ 0, & \text{if } x = 0, \\ f(x), & \text{if } 0 < x \le L. \end{cases}$$

We can also extend f to be an even function $f_e : [-L, L] \to \mathbb{C}$ by defining

$$f_e(x) = \begin{cases} f(-x), & \text{if } -L \le x < 0, \\ f(x), & \text{if } 0 \le x \le L. \end{cases}$$

The Fourier series of $f_o : [-L, L] \to \mathbb{C}$ is called the Fourier sine series of $f : [0, L] \to \mathbb{C}$. The Fourier series of $f_e : [-L, L] \to \mathbb{C}$ is called the Fourier cosine series of $f : [0, L] \to \mathbb{C}$.

Exercises 7.2

Question 1

Consider the function $f : [-\pi, \pi] \to \mathbb{R}$, $f(x) = x^3 - \pi^2 x$.

(a) Find the Fourier series of f.

(b) Use the Fourier series to find the sum

$$\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{(2k-1)^3} = 1 - \frac{1}{3^3} + \frac{1}{5^3} - \frac{1}{7^3} + \cdots .$$

Question 2

Let $f : [-\pi, \pi]$ be the function defined as

$$f(x) = \begin{cases} x + \pi, & \text{if } -\pi \leq x < 0, \\ x - \pi, & \text{if } 0 \leq x \leq \pi. \end{cases}$$

(a) Find the Fourier series of f.

(b) Study the pointwise convergence of the Fourier series.

Question 3

Study the pointwise convergence of the Fourier series of the function $f : [-\pi, \pi] \to \mathbb{R}$, $f(x) = |x|$ obtained in Exercises 7.1.

Question 4

Study the pointwise convergence of the Fourier series of the function $f : [-\pi, \pi] \to \mathbb{R}$, $f(x) = x^2$ obtained in Exercises 7.1.

Question 5

Study the pointwise convergence of the Fourier series of the function $f : [0, 2\pi] \to \mathbb{R}$, $f(x) = x^2$ obtained in Exercises 7.1.

7.3 The L^2 Convergence of a Fourier Series

In this section, we consider the L^2-convergence of a Fourier series. We first define L^2-converges for a sequence of Riemann integrable functions.

Definition 7.10 L^2-Convergence

Let $I = [a, b]$ be an interval in $\mathbb{R}$, and let $\{f_n : I \to \mathbb{C}\}$ be a sequence of Riemann integrable functions. We say that $\{f_n : I \to \mathbb{C}\}$ converges in L^2 to a function $g : I \to \mathbb{C}$ in $\mathcal{R}(I, \mathbb{C})$ if

$$\lim_{n\to\infty} \|f_n - g\| = 0.$$

In the vector space $\mathcal{R}(I, \mathbb{C})$, we have nonzero functions $h : I \to \mathbb{C}$ which has zero norm. Hence, if $\{f_n : I \to \mathbb{C}\}$ converges in L^2 to a function $g : I \to \mathbb{C}$, the function g is not unique. Nevertheless, we have the following.

Theorem 7.15

Let $I = [a, b]$ be an interval in $\mathbb{R}$, and let $\{f_n : I \to \mathbb{C}\}$ be a sequence of functions in $\mathcal{R}(I, \mathbb{C})$ that converges in L^2 to the two functions $g_1 : I \to \mathbb{C}$ and $g_2 : I \to \mathbb{C}$ in $\mathcal{R}(I, \mathbb{C})$, then g_1 and g_2 are L^2-equivalent.

Proof

By triangle inequality,

$$\|g_1 - g_2\| \le \|f_n - g_1\| + \|f_n - g_2\|.$$

Since

$$\lim_{n\to\infty} \|f_n - g_1\| = 0 \quad \text{and} \quad \lim_{n\to\infty} \|f_n - g_2\| = 0,$$

we find that $\|g_1 - g_2\| = 0$. Thus, g_1 and g_2 are L^2-equivalent.

Example 7.13

Consider the sequence of functions $\{f_n : [0,1] \to \mathbb{R}\}$ defined as

$$f_n(x) = \begin{cases} 1, & \text{if } x = \dfrac{m}{n} \text{ for some integer } m, \\ 0, & \text{otherwise.} \end{cases}$$

Since $f_n : [0,1] \to \mathbb{R}$ is a function that is nonzero only for finitely many points, we find that $\|f_n\| = 0$. This implies that $\{f_n : I \to \mathbb{R}\}$ converges in L^2 to the function $f_0 : I \to \mathbb{R}$ that is identically zero. However, $\{f_n : I \to \mathbb{R}\}$ does not converge pointwise. Take for example the point $x_0 = 1/2$. Then $f_n(x_0) = 1$ if n is even and $f_n(x_0) = 0$ if n is odd. Hence, the sequence $\{f_n(x_0)\}$ does not converge. In other words, for sequences of functions, pointwise convergence and L^2-convergence are different.

The Fourier series of a Riemann integrable function $f : I \to \mathbb{C}$ converges in L^2 to the function $f : I \to \mathbb{C}$ if

$$\lim_{n\to\infty} \|s_n - f\| = 0.$$

Here $s_n(x)$ is the n^{th}-partial sum of the Fourier series. The main theorem we want to prove in this section is the Fourier series of any Riemann integrable function $f : I \to \mathbb{C}$ converges in L^2 to f itself. We start with the following theorem which asserts that a Riemann integrable function $f : [a,b] \to \mathbb{C}$ can be approximated in L^2 by step functions.

Theorem 7.16

Let $f : [a,b] \to \mathbb{C}$ be a Riemann integrable function. For every $\varepsilon > 0$, there is a step function $g : [a,b] \to \mathbb{C}$ such that

$$\|f - g\| < \varepsilon.$$

Proof

As in the proof of Theorem 7.9, it is sufficient to consider the case where the function f is real-valued. Since $f : [a, b] \to \mathbb{R}$ is Riemann integrable, it is bounded. Therefore, there exists $M > 0$ such that

$$|f(x)| \leq M \qquad \text{for all } x \in [a, b].$$

Given $\varepsilon > 0$, Theorem 7.9 says that there is a step function $g : [a, b] \to \mathbb{R}$ such that

$$\int_a^b |f(x) - g(x)|dx < \frac{\varepsilon^2}{2M}.$$

By the construction of g given in the proof of Theorem 7.9, we find that $|g(x)| \leq M$ for all $x \in [a, b]$. Therefore,

$$(f(x) - g(x))^2 \leq |f(x) - g(x)||f(x) + g(x)| \leq 2M|f(x) - g(x)|.$$

This implies that

$$\|f - g\|^2 = \int_a^b (f(x) - g(x))^2 dx \leq 2M \int_a^b |f(x) - g(x)|dx < \varepsilon^2.$$

Hence, $\|f - g\| < \varepsilon$.

Theorem 7.17

Let $I = [-\pi, \pi]$. Given a Riemann integrable function $f : I \to \mathbb{C}$, let $\sum_{k=-\infty}^{\infty} c_k e^{ikx}$ be its Fourier series, and let $s_n(x) = \sum_{k=-n}^{n} c_k e^{ikx}$ be the n^{th}-partial sum. We have the followings.

(a) For each $n \geq 0$, $\|s_n\|^2 = 2\pi \sum_{k=-n}^{n} |c_k|^2$.

(b) For each $n \geq 0$, we have the Bessel's inequality $\|s_n\| \leq \|f\|$.

(c) The Fourier series converges in L^2 to f if and only if

$$\lim_{n\to\infty} \|s_n\|^2 = \|f\|^2.$$

Proof

For $n \in \mathbb{Z}$, let $\phi_n : \mathbb{R} \to \mathbb{C}$ be the function $\phi_n(x) = e^{inx}$. Then $\mathcal{S} = \{\phi_n \,|\, n \in \mathbb{Z}\}$ is an orthogonal system of functions in $\mathcal{R}(I, \mathbb{C})$, and $\|\phi_n\| = \sqrt{2\pi}$ for all $n \in \mathbb{Z}$. For $n \geq 0$, the set $\mathcal{S}_n = \{\phi_k \,|\, -n \leq k \leq n\}$ spans the subspace W_n, and

$$\sum_{k=-n}^{n} c_k \phi_k(x) = s_n(x) = \left(\text{proj}_{W_n} f\right)(x).$$

Since $\mathcal{S}_n$ is an orthogonal system, the generalized Pythagoras theorem says that

$$\|s_n\|^2 = \sum_{k=-n}^{n} |c_k|^2 \|\phi_k\|^2 = 2\pi \sum_{k=-n}^{n} |c_k|^2.$$

This proves part (a).

For part (b), recall that $f - s_n$ is orthogonal to s_n. By generalized Pythagoras theorem again,

$$\|f\|^2 = \|s_n + (f - s_n)\|^2 = \|s_n\|^2 + \|f - s_n\|^2 \geq \|s_n\|^2.$$

Hence, we find that $\|s_n\| \leq \|f\|$. This proves part (b).

Part (c) follows from

$$\|f\|^2 = \|s_n\|^2 + \|f - s_n\|^2.$$

The Fourier series converges in L^2 to f if and only if $\lim_{n\to\infty} \|s_n - f\| = 0$, if and only if

$$\lim_{n\to\infty} \|s_n\|^2 = \|f\|^2.$$

Remark 7.9

Part (b) of Theorem 7.17 says for the trigonometric series $\sum_{k=-\infty}^{\infty} c_k e^{ikx}$ to be the Fourier series of a Riemann integrable function, it is necessary that the series $\sum_{k=-\infty}^{\infty} |c_k|^2$ is convergent.

Now we will prove that the Fourier series of a special type of step functions converges in L^2 to the function itself.

Theorem 7.18

Let a and b be two numbers satisfying $-\pi \leq a < b \leq \pi$, and let $g : [-\pi, \pi] \to \mathbb{R}$ be the function defined as

$$g(x) = \begin{cases} 1, & \text{if } a \leq x \leq b, \\ 0, & \text{otherwise.} \end{cases}$$

The Fourier series of g converges in L^2 to the function g.

Proof

By Theorem 7.17, it is sufficient to show that

$$\lim_{n\to\infty} \|s_n\|^2 = \|g\|^2.$$

Now,

$$\|g\|^2 = \int_{-\pi}^{\pi} |g(x)|^2 dx = \int_a^b dx = b - a.$$

In Example 7.7, we have seen that the Fourier coefficients of g is

$$c_0 = \frac{b-a}{2\pi}, \quad \text{and} \quad c_k = \frac{e^{-ikb} - e^{-ika}}{-2\pi ik} \quad \text{when } k \neq 0.$$

By part (a) of Theorem 7.17,

$$\begin{aligned} \|s_n\|^2 &= 2\pi \left(|c_0|^2 + \sum_{k=1}^{n} (|c_k|^2 + |c_{-k}|^2) \right) \\ &= \frac{(b-a)^2}{2\pi} + 4\pi \sum_{k=1}^{n} \frac{(e^{ikb} - e^{ika})(e^{-ikb} - e^{-ika})}{4\pi^2 k^2} \\ &= \frac{(b-a)^2}{2\pi} + \frac{2}{\pi} \sum_{k=1}^{n} \frac{1 - \cos k(b-a)}{k^2}. \end{aligned}$$

By Example 7.11,

$$\frac{2}{3}\pi^2 - 4\sum_{k=1}^{\infty}\frac{\cos kx}{k^2} = x(2\pi - x) \qquad \text{for all } x \in [0, 2\pi],$$

and

$$\sum_{k=1}^{\infty}\frac{1}{k^2} = \frac{\pi^2}{6}.$$

Therefore,

$$\begin{aligned}\lim_{n\to\infty}\|s_n\|^2 &= \frac{(b-a)^2}{2\pi} + \frac{2}{\pi}\sum_{k=1}^{\infty}\frac{1}{k^2} - \frac{2}{\pi}\sum_{k=1}^{\infty}\frac{\cos k(b-a)}{k^2} \\ &= \frac{(b-a)^2}{2\pi} + \frac{\pi}{3} - \frac{1}{2\pi}\left(\frac{2}{3}\pi^2 - 2\pi(b-a) + (b-a)^2\right) \\ &= b - a = \|g\|^2.\end{aligned}$$

This proves that the Fourier series of g converges in L^2 to g.

Now we can prove our main theorem.

Theorem 7.19 **L^2 Convergence of Fourier Series**

Let $I = [-\pi, \pi]$ and let $f : I \to \mathbb{C}$ be a Riemann integrable function. Then the Fourier series of f converges in L^2 to f itself.

Proof

Since we will be dealing with more than one functions here, we use $s_n(f)$ to denote the n^{th}-partial sum of the Fourier series of f.

We will show that given $\varepsilon > 0$, there exists a positive integer N such that for all $n \geq N$,

$$\|s_n(f) - f\| < \varepsilon.$$

Fixed $\varepsilon > 0$. Theorem 7.16 says that there is step function $g : [-\pi, \pi] \to \mathbb{R}$ such that

$$\|f - g\| < \frac{\varepsilon}{3}.$$

Let $P = \{x_0, x_1, \ldots, x_l\}$ be the partition of $[-\pi, \pi]$ such that for each $1 \leq j \leq l$, g is constant on (x_{j-1}, x_j). Define $g_j : [-\pi, \pi] \to \mathbb{C}$ by

$$g_j(x) = \begin{cases} g(x), & \text{if } x_{j-1} < x < x_j, \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$g(x) = g_1(x) + \cdots + g_l(x) \qquad \text{for all } x \in [-\pi, \pi] \setminus P.$$

Since Riemann integrals are not affected by function values at finitely many points, it follows that for each $n \geq 0$,

$$s_n(g) = \sum_{j=1}^{l} s_n(g_j).$$

By Theorem 7.17,

$$\|s_n(f) - s_n(g)\| = \|s_n(f - g)\| \leq \|f - g\| < \frac{\varepsilon}{3}.$$

By triangle inequality,

$$\|g - s_n(g)\| = \left\| \sum_{j=1}^{l} (g_j - s_n(g_j)) \right\| \leq \sum_{j=1}^{l} \|g_j - s_n(g_j)\|.$$

By Theorem 7.18, $\lim_{n\to\infty} \|g_j - s_n(g_j)\| = 0$ for $1 \leq j \leq l$. Therefore,

$$\lim_{n\to\infty} \|g - s_n(g)\| = 0.$$

This implies that there is a positive integer N such that

$$\|g - s_n(g)\| < \frac{\varepsilon}{3} \qquad \text{for all } n \geq N.$$

It follows that for all $n \geq N$,

$$\|f - s_n(f)\| \leq \|f - g\| + \|g - s_n(g)\| + \|s_n(f) - s_n(g)\| < \varepsilon.$$

This completes the proof.

A consequence of Theorem 7.19 is the following.

Theorem 7.20 Parseval's Identity I

Let $f : [-\pi, \pi] \to \mathbb{C}$ be a Riemann integrable function, and let

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-ikx} dx, \qquad k \in \mathbb{Z}$$

be its Fourier coefficients. Then

$$\int_{-\pi}^{\pi} |f(x)|^2 dx = \|f\|^2 = 2\pi \sum_{k=-\infty}^{\infty} |c_k|^2.$$

Proof

In Theorem 7.19, we have shown that the Fourier series of f converges in L^2 to f. By Theorem 7.17, this means that

$$\lim_{n\to\infty} \|s_n\|^2 = \|f\|^2.$$

Since

$$\|s_n\|^2 = 2\pi \sum_{k=-n}^{n} |c_k|^2,$$

the Parseval's identity follows.

Corollary 7.21 Parseval's Identity II

Let $f : [-\pi, \pi] \to \mathbb{R}$ be a real-valued Riemann integrable function, and let

$$a_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos kx dx, \qquad k \geq 0,$$

$$b_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin kx dx, \qquad k \geq 1$$

be its Fourier coefficients. Then

$$\int_{-\pi}^{\pi} f(x)^2 dx = \|f\|^2 = \pi \left\{ \frac{a_0^2}{2} + \sum_{k=1}^{\infty} \left(a_k^2 + b_k^2\right) \right\}.$$

Proof

This can be proved using $c_0 = \dfrac{a_0}{2} \in \mathbb{R}$, and when $k \geq 1$,

$$c_k = \frac{a_k - ib_k}{2}, \qquad c_{-k} = \frac{a_k + ib_k}{2},$$

and a_k and b_k are real numbers.

Let us look at a few examples.

Example 7.14

In Example 7.4, we have seen that the Fourier series of the function $f : [-\pi, \pi] \to \mathbb{R}$, $f(x) = x$ is

$$F(x) = \sum_{k=1}^{\infty} 2(-1)^{k-1} \frac{\sin kx}{k}.$$

Using Parseval's identity, we deduce that

$$\pi \sum_{k=1}^{\infty} \frac{4}{k^2} = \int_{-\pi}^{\pi} x^2 dx = \frac{2\pi^3}{3}.$$

This gives

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6},$$

an identity we have obtained before.

Example 7.15

In Example 7.5, we have seen that the Fourier series of the function $f : [0, 2\pi] \to \mathbb{R}$, $f(x) = x(2\pi - x)$ is

$$F(x) = \frac{2}{3}\pi^2 - 4 \sum_{k=1}^{\infty} \frac{\cos kx}{k^2}.$$

Using Parseval's identity, we have

$$\begin{aligned}
&2\pi \times \frac{4}{9}\pi^4 + 16\pi \sum_{k=1}^{\infty} \frac{1}{k^4} \\
&= \int_0^{2\pi} x^2(2\pi - x)^2 dx = \int_0^{2\pi} \left(4\pi^2 x^2 - 4\pi x^3 + x^4\right) dx \\
&= \left[\frac{4\pi^2 x^3}{3} - \pi x^4 + \frac{x^5}{5}\right]_0^{2\pi} \\
&= \left(\frac{32}{3} - 16 + \frac{32}{5}\right)\pi^5 = \frac{16}{15}\pi^5.
\end{aligned}$$

Therefore,

$$\sum_{k=1}^{\infty} \frac{1}{k^4} = \frac{\pi^4}{15} - \frac{\pi^4}{18} = \frac{\pi^4}{90}.$$

From the Parseval's identity, we can also obtain the following.

Theorem 7.22

Let $I = [-\pi, \pi]$. Given that $f : I \to \mathbb{C}$ and $g : I \to \mathbb{C}$ are Riemann integrable functions, let

$$c_k(f) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-ikx}dx, \qquad c_k(g) = \frac{1}{2\pi}\int_{-\pi}^{\pi} g(x)e^{-ikx}dx$$

be their Fourier coefficients. Then

$$\int_{-\pi}^{\pi} f(x)\overline{g(x)}dx = \langle f, g\rangle = 2\pi \sum_{k=-\infty}^{\infty} c_k(f)\overline{c_k(g)}.$$

Proof

This follows from Theorem 7.20 and the polarization formula

$$\begin{aligned}
\langle f, g\rangle = \frac{1}{4}\big(&\langle f+g, f+g\rangle - \langle f-g, f-g\rangle \\
&+i\langle f+ig, f+ig\rangle - i\langle f-ig, f-ig\rangle\big).
\end{aligned}$$

We can use Theorem 7.22 to prove that Fourier series can be integrated term

by term.

Theorem 7.23 Term-by-Term Integration of Fourier Series

Let $I = [-\pi, \pi]$. Given that $f : I \to \mathbb{C}$ is a Riemann integrable function, let $\sum_{k=-\infty}^{\infty} c_k e^{ikx}$ be its Fourier series. On any compact interval $J = [a, b]$ that is contained in I, we can integrate term by term and obtain

$$\int_a^b f(x)dx = \sum_{k=-\infty}^{\infty} c_k \int_a^b e^{ikx}dx.$$

Proof

Let $g : [-\pi, \pi] \to \mathbb{R}$ be the function defined as

$$g(x) = \begin{cases} 1, & \text{if } a \leq x \leq b, \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$c_k(g) = \frac{1}{2\pi} \int_a^b e^{-ikx}dx.$$

Using Theorem 7.22, we find that

$$\begin{aligned} \int_a^b f(x)dx &= \int_{-\pi}^{\pi} f(x)\overline{g(x)}dx \\ &= 2\pi \sum_{k=-\infty}^{\infty} c_k(f)\overline{c_k(g)} \\ &= \sum_{k=\infty}^{\infty} c_k(f) \int_a^b e^{ikx}dx. \end{aligned}$$

This proves the assertion.

Remark 7.10

Theorem 7.23 is remarkable since we do not require the Fourier series of f to converge uniformly.

Example 7.16

In Example 7.4, we have seen that the Fourier series of the function $f : [-\pi, \pi] \to \mathbb{R}$, $f(x) = x$ is

$$F(x) = \sum_{k=1}^{\infty} 2(-1)^{k-1} \frac{\sin kx}{k}.$$

For $x \in [-\pi, \pi]$, term-by-term integration gives

$$\int_0^x t dt = \sum_{k=1}^{\infty} (-1)^{k-1} \frac{2}{k} \int_0^x \sin kt dt.$$

This implies that

$$\frac{x^2}{2} = \sum_{k=1}^{\infty} (-1)^{k-1} \frac{2(1 - \cos kx)}{k^2}.$$

Since

$$\sum_{k=1}^{\infty} (-1)^{k-1} \frac{1}{k^2} = \frac{\pi^2}{12},$$

we find that

$$x^2 = \frac{\pi^2}{3} + \sum_{k=1}^{\infty} (-1)^k \frac{4}{k^2} \cos kx.$$

Theorem 7.24 General Term-by-Term Integration of Fourier Series

Let $I = [-\pi, \pi]$. Given that $f : I \to \mathbb{C}$ is a Riemann integrable function, let $\sum_{k=-\infty}^{\infty} c_k e^{ikx}$ be its Fourier series. Let $g : I \to \mathbb{C}$ be any other Riemann integrable function. On any compact interval $J = [a, b]$ that is contained in I, we have

$$\int_a^b f(x)g(x)dx = \sum_{k=-\infty}^{\infty} c_k \int_a^b g(x)e^{ikx} dx.$$

Proof

Let $h : [-\pi, \pi] \to \mathbb{R}$ be the function defined as

$$h(x) = \begin{cases} \overline{g(x)}, & \text{if } a \leq x \leq b, \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$c_k(h) = \frac{1}{2\pi} \int_a^b \overline{g(x)} e^{-ikx} dx.$$

Hence,

$$\overline{c_k(h)} = \frac{1}{2\pi} \int_a^b g(x) e^{ikx} dx.$$

Using Theorem 7.22, we find that

$$\begin{aligned} \int_a^b f(x)g(x)dx &= \int_{-\pi}^{\pi} f(x)\overline{h(x)}dx \\ &= 2\pi \sum_{k=-\infty}^{\infty} c_k(f)\overline{c_k(h)} \\ &= \sum_{k=-\infty}^{\infty} c_k(f) \int_a^b g(x)e^{ikx}dx. \end{aligned}$$

This proves the assertion.

Exercises 7.3

Question 1

Consider the function $f : [-\pi, \pi] \to \mathbb{R}$, $f(x) = x^3 - \pi^2 x$ whose Fourier series has been obtained in Excerses 7.2. Use Parseval's identity to find the sum

$$\sum_{k=1}^{\infty} \frac{1}{k^6} = 1 + \frac{1}{2^6} + \frac{1}{3^6} + \frac{1}{4^6} + \cdots .$$

Question 2

Use term by term integration to find the Fourier series of the function $f : [-\pi, \pi] \to \mathbb{R}$, $f(x) = x^3$.

7.4 The Uniform Convergence of a Trigonometric Series

In this section, we consider uniform convergence of a trigonometric series. In volume I, we have studied uniform convergence. If a series of continuous functions converges uniformly, then it represents a continuous function, and the series can be integrated term-by-term.

Applying to trigonometric series, we have the following.

Theorem 7.25

If the trigonometric series $\sum_{k=-\infty}^{\infty} c_k e^{ikx}$ converges uniformly, it defines a continuous 2π-periodic function $F : \mathbb{R} \to \mathbb{C}$. The Fourier series of $F : [-\pi, \pi] \to \mathbb{C}$ is the series $\sum_{k=-\infty}^{\infty} c_k e^{ikx}$ itself.

Proof

For $k \in \mathbb{Z}$, the function $\phi_k(x) = e^{ikx}$ is a 2π-periodic continuous function. The assertion that the trigonometric series $\sum_{k=-\infty}^{\infty} c_k e^{ikx}$ defines a continuous 2π-periodic function $F : \mathbb{R} \to \mathbb{C}$ follows from the uniform convergence. Since each $\phi_k(z)$, $k \in \mathbb{Z}$ is Riemann integrable, and the series $\sum_{k=-\infty}^{\infty} c_k e^{ikx}$ converges uniformly, we can integrate term by term to find that

$$c_k(F) = \frac{1}{2\pi} \int_{-\pi}^{\pi} F(x) e^{-ikx} dx = \frac{1}{2\pi} \sum_{l=-\infty}^{\infty} c_l \int_{-\pi}^{\pi} e^{i(l-k)x} dx = c_k.$$

This proves that the Fourier series of $F : [-\pi, \pi] \to \mathbb{C}$ is the series $\sum_{k=-\infty}^{\infty} c_k e^{ikx}$ itself.

We would like to have a sufficient condition for a trigonometric series

$$\sum_{k=-\infty}^{\infty} c_k e^{ikx}$$

to converge uniformly. In volume I, we have discussed Weierstrass M-test for

real-valued functions. This can be generalized to complex-valued functions in a straightforward way. Let $\{f_n : D \to \mathbb{C}\}$ be a sequence of functions defined on a subset D of $\mathbb{R}$. Assume that for each $n \in \mathbb{Z}^+$, there is a constant M_n so that

$$|f_n(x)| \leq M_n \qquad \text{for all } x \in D.$$

The Weiertrass M-test states that if $\displaystyle\sum_{n=1}^{\infty} M_n$ is convergent, then the series of functions $\displaystyle\sum_{n=1}^{\infty} f_n(x)$ converges absolutely and uniformly.

Using Weierstrass M-test, we can deduce the following.

Theorem 7.26

If the series $\displaystyle\sum_{k=-\infty}^{\infty} |c_k|$ is convergent, then the trigonometric series $\displaystyle\sum_{k=-\infty}^{\infty} c_k e^{ikx}$ converges absolutely and uniformly, and it defines a continuous function $F : \mathbb{R} \to \mathbb{C}$ whose Fourier series is itself.

Proof

For any $k \in \mathbb{Z}$,

$$|c_k e^{ikx}| \leq |c_k| \qquad \text{for all } x \in \mathbb{R}.$$

By Weierstrass M-test, the series $\displaystyle\sum_{k=-\infty}^{\infty} c_k e^{ikx}$ converges absolutely and uniformly. The rest follows from Theorem 7.25.

Example 7.17

Consider the series

$$\sum_{k=1}^{\infty} \frac{\cos kx}{k^{\frac{3}{2}}}.$$

Since $\sum_{k=1}^{\infty} \frac{1}{k^{\frac{3}{2}}}$ is a p-series with $p = \frac{3}{2} > 1$, it is convergent. Hence, the function $F : \mathbb{R} \to \mathbb{R}$ defined as

$$F(x) = \sum_{k=1}^{\infty} \frac{\cos kx}{k^{\frac{3}{2}}}$$

is a continuous function whose Fourier series is itself.

Remark 7.11

By Remark 7.9, in order for the series $\sum_{k=-\infty}^{\infty} c_k e^{ikx}$ to be the Fourier series of a Riemann integrable function, it is necessary that the series $\sum_{k=-\infty}^{\infty} |c_k|^2$ is convergent. However, the convergence of $\sum_{k=-\infty}^{\infty} |c_k|^2$ does not imply the convergence of $\sum_{k=-\infty}^{\infty} |c_k|$.

Theorem 7.25 gives a criterion for a function defined as a trigonometric series to have Fourier series that is equal to itself. However, we will usually start by a Riemann integrable function.

Theorem 7.27

Let $I = [-\pi, \pi]$, and let $f : I \to \mathbb{C}$ be a Riemann integrable function. If the Fourier series $\sum_{k=-\infty}^{\infty} c_k e^{ikx}$ of $f : I \to \mathbb{C}$ converges uniformly, it defines a continuous 2π-periodic function $F : \mathbb{R} \to \mathbb{C}$ whose restriction to I is L^2 equivalent to $f : I \to \mathbb{C}$. If the periodic extension $\widetilde{f} : \mathbb{R} \to \mathbb{C}$ of f is continuous at x_0, then $F(x_0) = \widetilde{f}(x_0)$.

Proof

The fact that $\sum_{k=-\infty}^{\infty} c_k e^{ikx}$ defines a continuous periodic function $F : \mathbb{R} \to \mathbb{C}$ has been asserted in Theorem 7.25. Since $\sum_{k=-\infty}^{\infty} c_k e^{ikx}$ is the Fourier series for both $f : I \to \mathbb{C}$ and $F : I \to \mathbb{C}$, Theorem 7.19 says that it converges in L^2 to $f : I \to \mathbb{C}$ and $F : I \to \mathbb{C}$. Theorem 7.15 then asserts that $F : I \to \mathbb{C}$ and $f : I \to \mathbb{C}$ are L^2 equivalent. The values of two L^2-equivalent functions agree at a point where both of them are continuous.

Example 7.18

In Example 7.5, we have seen that the Fourier series of the function $f : [0, 2\pi] \to \mathbb{R}$, $f(x) = x(2\pi - x)$ is

$$F(x) = \frac{2}{3}\pi^2 - 4\sum_{k=1}^{\infty} \frac{\cos kx}{k^2}.$$

Since the series $\sum_{k=1}^{\infty} \frac{1}{k^2}$ is convergent, and $f : [0, 2\pi] \to \mathbb{R}$ is continuous with $f(0) = f(2\pi) = 0$, Theorem 7.26 and Theorem 7.27 imply that the Fourier series $F(x)$ converges uniformly to the periodic extension of the function $f(x)$.

Example 7.19

In Example 7.4, we have seen that the Fourier series of the function $f : [-\pi, \pi] \to \mathbb{R}$, $f(x) = x$ is

$$F(x) = \sum_{k=1}^{\infty} 2(-1)^{k-1} \frac{\sin kx}{k}.$$

Since the harmonic series $\sum_{k=1}^{\infty} \frac{1}{k}$ is divergent, we cannot apply Theorem 7.26. However, we can argue that the Fourier series does not converge uniformly in the following way.

In Example 7.10, we have used Dirichlet's theorem to conclude that $F(x)$ converges to $f(x)$ for any $x \in (-\pi, \pi)$. It is obvious that $F(\pi) = 0$.
Now

$$\lim_{x\to\pi^-} F(x) = \lim_{x\to\pi^-} f(x) = \lim_{x\to\pi^-} x = \pi \neq F(\pi).$$

This shows that F is not continuous at $x = \pi$. Hence, the convergence of the Fourier series is not uniform.

Remark 7.12

Theorem 7.25 and Theorem 7.27 can be regarded as uniquessness of Fourier series.

Next we want to consider term-by-term differentiation.

Example 7.20

Let us consider again the function $f : [-\pi, \pi] \to \mathbb{R}$, $f(x) = x$, whose Fourier series is given by

$$F(x) = \sum_{k=1}^{\infty} 2(-1)^{k-1} \frac{\sin kx}{k}.$$

Since $f : (-\pi, \pi) \to \mathbb{R}$ is continuously differentiable with $f'(x) = 1$, the extension function $\widetilde{f} : \mathbb{R} \to \mathbb{R}$ is continuously differentiable at all the points where $x \notin (2n+1)\mathbb{Z}$, and

$$\widetilde{f}'(x) = 1, \quad \text{when } x \notin (2n+1)\mathbb{Z}.$$

Hence, for the function $f' : [-\pi, \pi] \to \mathbb{R}$, regardless of how it is defined at $\pm\pi$, its Fourier series is the constant $G(x) = 1$. However, if we differentiate the Fourier series of $f : [-\pi, \pi] \to \mathbb{R}$ term-by-term, we obtain the series

$$\sum_{k=1}^{\infty} 2(-1)^{k-1} \cos kx.$$

As $|a_k| = 2$ for all $k \in \mathbb{Z}$, $\sum_{k=1}^{\infty} |a_k|^2$ is divergent. Thus, the series

$$\sum_{k=1}^{\infty} 2(-1)^{k-1} \cos kx.$$

cannot be the Fourier series of any Riemann-integrable function.

Example 7.20 shows that term-by-term differentiation can fail even though the function $f : [-\pi, \pi] \to \mathbb{R}$ and its derivative are strongly piecewise differentiable.

Theorem 7.28

Let $I = [-\pi, \pi]$ and let $f : I \to \mathbb{R}$ be a Riemann integrable function with Fourier series $\sum_{k=-\infty}^{\infty} c_k e^{ikx}$. If the series $\sum_{k=-\infty}^{\infty} c_k$ is convergent, and the series $\sum_{k=-\infty}^{\infty} kc_k e^{ikx}$ converges uniformly, then $f : I \to \mathbb{R}$ is L^2-equivalent to the continuously differentiable function $F : I \to \mathbb{R}$, $F(x) = \sum_{k=-\infty}^{\infty} c_k e^{ikx}$. Moreover, the Fourier series of $F' : I \to \mathbb{R}$ is

$$\sum_{k=-\infty}^{\infty} ikc_k e^{ikx},$$

which converges to $F'(x)$ for all $x \in I$.

Proof

By Theorem 7.25, the uniformly convergent series

$$\sum_{k=-\infty}^{\infty} ikc_k e^{ikx}$$

defines a continuous function $G : I \to \mathbb{C}$, whose Fourier series is itself.

Let

$$H(x) = \int_0^x G(x)dx + \sum_{k=-\infty}^{\infty} c_k.$$

By fundamental theorem of calculus, H is differentiable and $H'(x) = G(x)$. Thus, $H(x)$ is continuously differentiable. Since the series

$$\sum_{k=-\infty}^{\infty} ikc_k e^{ikx}$$

converges uniformly, we can do term-by-term integration to obtain

$$H(x) = \sum_{k=-\infty}^{\infty} ikc_k \int_0^x e^{ikt}dt + \sum_{k=-\infty}^{\infty} c_k = \sum_{k=-\infty}^{\infty} c_k e^{ikx} = F(x).$$

Hence, the function $F : I \to \mathbb{C}$,

$$F(x) = \sum_{k=-\infty}^{\infty} c_k e^{ikx}$$

is continuously differentiable, with derivative

$$F'(x) = H'(x) = G(x) = \sum_{k=-\infty}^{\infty} ikc_k e^{ikx}.$$

This completes the proof.

Let us look at an example.

Example 7.21

Consider the trigonometric series

$$\sum_{k=1}^{\infty} \frac{\sin kx}{k^3} \quad \text{and} \quad \sum_{k=1}^{\infty} \frac{\cos kx}{k^2}.$$

Since the series $\sum_{k=1}^{\infty} \frac{1}{k^3}$ and $\sum_{k=1}^{\infty} \frac{1}{k^2}$ are convergent, both the series

$$\sum_{k=1}^{\infty} \frac{\sin kx}{k^3} \qquad \text{and} \qquad \sum_{k=1}^{\infty} \frac{\cos kx}{k^2}$$

converge uniformly, and so they define continuous functions. Let

$$F(x) = \sum_{k=1}^{\infty} \frac{\sin kx}{k^3} \qquad \text{and} \qquad G(x) = \sum_{k=1}^{\infty} \frac{\cos kx}{k^2}.$$

Since

$$\frac{d}{dx}\frac{\sin kx}{k^3} = \frac{\cos kx}{k^2} \qquad \text{for all } k \in \mathbb{Z}^+,$$

F is continuously differentiable and $F'(x) = G(x)$.

At the end of this section, we give a brief discussion about the Cesàro mean of a Fourier series. As an application, we give another proof of the Weiestrass approximation theorem.

Definition 7.11 Cesàro Mean of a Fourier Series

Given a Riemann integrable function $f : [-\pi, \pi] \to \mathbb{C}$, let its Fourier series be

$$\sum_{k=-\infty}^{\infty} c_k e^{ikx}, \qquad \text{where } c_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-ikx}dx.$$

For $n \geq 1$, the n^{th} Cesàro mean of the Fourier series is

$$\sigma_n(x) = \frac{s_0(x) + s_1(x) + \cdots + s_{n-1}(x)}{n},$$

where

$$s_n(x) = \sum_{k=-n}^{n} c_k e^{ikx}$$

is the n^{th}-partial sum of the Fourier series.

Proposition 7.29

Let $I = [-\pi, \pi]$, and let $f : I \to \mathbb{C}$ be a Riemann integrable function. The n^{th} Cesàro mean $\sigma_n(x)$ of the Fourier series of f has an integral representation given by

$$\sigma_n(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \mathcal{F}_n(x-t) dt,$$

where $\mathcal{F}_n : \mathbb{R} \to \mathbb{R}$ is the kernel

$$\mathcal{F}_n(x) = \begin{cases} n, & \text{if } x \in 2\pi\mathbb{Z}, \\ \dfrac{\sin^2 \frac{nx}{2}}{n \sin^2 \frac{x}{2}}, & \text{otherwise.} \end{cases}$$

Proof

By Proposition 7.8,

$$s_n(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) D_n(x-t) dt,$$

where $D_n(t)$ is the Dirichlet kernel. This gives

$$\sigma_n(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \mathcal{F}_n(x-t) dt,$$

where

$$\mathcal{F}_n(x) = \frac{D_0(x) + D_1(x) + \cdots + D_{n-1}(x)}{n}.$$

Using the fact that

$$D_n(x) = \sum_{k=-n}^{n} e^{ikx} = \begin{cases} 2n+1, & \text{if } x \in 2\pi\mathbb{Z}, \\ \dfrac{\sin\left(n + \frac{1}{2}\right) x}{\sin \frac{x}{2}}, & \text{otherwise,} \end{cases}$$

we find that when $x \in 2\pi\mathbb{Z}$,

$$\mathcal{F}_n(x) = \frac{1 + 3 + \ldots + (2n-1)}{n} = n.$$

When $x \notin 2\pi\mathbb{Z}$,

$$\begin{aligned}
\mathcal{F}_n(x) &= \frac{1}{n \sin \frac{x}{2}} \sum_{k=0}^{n-1} \sin\left(k + \frac{1}{2}\right) x \\
&= \frac{1}{n \sin \frac{x}{2}} \operatorname{Im} \left\{ \sum_{k=0}^{n-1} \exp\left(i \left[k + \frac{1}{2}\right] x\right) \right\}. \\
&= \frac{1}{n \sin \frac{x}{2}} \operatorname{Im} \left\{ e^{\frac{ix}{2}} \frac{e^{inx} - 1}{e^{ix} - 1} \right\} \\
&= \frac{1}{n \sin \frac{x}{2}} \operatorname{Im} \left\{ \frac{e^{inx} - 1}{2i \sin \frac{x}{2}} \right\} \\
&= \frac{1 - \cos nx}{2n \sin^2 \frac{x}{2}} \\
&= \frac{\sin^2 \frac{nx}{2}}{n \sin^2 \frac{x}{2}}.
\end{aligned}$$

This completes the proof.

Definition 7.12

For $n \geq 1$, the Fejér kernel $\mathcal{F}_n : \mathbb{R} \to \mathbb{R}$ is the kernel given by

$$\mathcal{F}_n(x) = \begin{cases} n, & \text{if } x \in 2\pi\mathbb{Z}, \\ \dfrac{\sin^2 \frac{nx}{2}}{n \sin^2 \frac{x}{2}}, & \text{otherwise.} \end{cases}$$

A good property about the Fejér kernel is $\mathcal{F}_n(t) \geq 0$ for all $t \in \mathbb{R}$.
Now we can prove the following theorem.

Theorem 7.30

Let $f : [-\pi, \pi] \to \mathbb{C}$ be a continuous function with $f(-\pi) = f(\pi)$, and let $\sigma_n(x)$ be the n^{th} Cesàro mean of the Fourier series of f. Then the sequence of functions $\{\sigma_n : [-\pi, \pi] \to \mathbb{C}\}$ converges uniformly to the function $f : [-\pi, \pi] \to \mathbb{C}$.

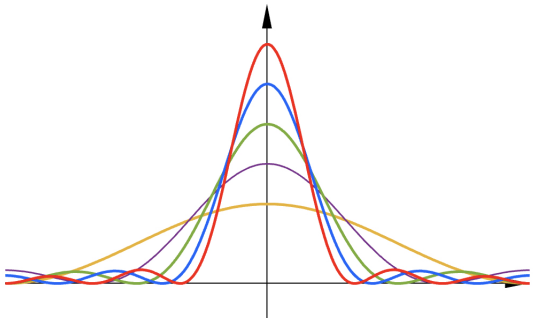


Figure 7.12: The Fejér kernels $\mathcal{F}_n : [-\pi, \pi] \to \mathbb{R}$ for $2 \leq n \leq 6$.

Proof

As in the proof of the Dirichlet's theorem, we find that

$$\sigma_n(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \widetilde{f}(x-t)\mathcal{F}_n(t)dt,$$

where $\widetilde{f} : \mathbb{R} \to \mathbb{C}$ be its 2π-periodic extension of f. Since

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} D_n(t)dt = 1 \qquad \text{for all } n \geq 0,$$

we have

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \mathcal{F}_n(t)dt = 1 \qquad \text{for all } n \geq 1.$$

It follows that for $x \in [-\pi, \pi]$,

$$\sigma_n(x) - f(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(\widetilde{f}(x-t) - \widetilde{f}(x)\right) \mathcal{F}_n(t)dt.$$

For $x \in [-\pi, \pi]$ and $t \in [-\pi, \pi]$, $x - t \in [-2\pi, 2\pi]$. Since $f : [-\pi, \pi] \to \mathbb{C}$ is a continuous function with $f(-\pi) = f(\pi)$, $\widetilde{f} : [-2\pi, 2\pi] \to \mathbb{C}$ is continuous. Hence, it is uniformly continuous.

Given $\varepsilon > 0$, there exists $\delta > 0$ such that if u and v are in $[-2\pi, 2\pi]$ and $|u - v| < \delta$, then

$$|\widetilde{f}(u) - \widetilde{f}(v)| < \frac{\varepsilon}{2}.$$

By continuity, if $|u - v| \leq \delta$, then

$$|\widetilde{f}(u) - \widetilde{f}(v)| \leq \frac{\varepsilon}{2}.$$

Being continuous on a compact interval, the function $\widetilde{f} : [-2\pi, 2\pi] \to \mathbb{C}$ is also bounded. Therefore, there exists $M > 0$ such that

$$|\widetilde{f}(x)| \leq M \qquad \text{for all } x \in [-2\pi, 2\pi].$$

This implies that for any $x \in [-\pi, \pi]$ and $t \in [-\pi, \pi]$,

$$\left|\widetilde{f}(x-t) - \widetilde{f}(x)\right| \leq 2M.$$

On the other hand, if $\delta \leq |t| \leq \pi$,

$$\sin^2 \frac{t}{2} \geq \sin^2 \frac{\delta}{2} > 0.$$

Therefore,

$$0 \leq \mathcal{F}_n(t) \leq \frac{1}{n \sin^2 \frac{\delta}{2}} \qquad \text{when } \delta \leq |t| \leq \pi.$$

Let N be a positive integer such that

$$N > \frac{4M}{\varepsilon \sin^2 \frac{\delta}{2}}.$$

For $n \geq N$ and $x \in [-\pi, \pi]$, we have

$$\begin{aligned} |\sigma_n(x) - f(x)| &\leq \frac{1}{2\pi} \int_{|t|\leq\delta} \left|\widetilde{f}(x-t) - \widetilde{f}(x)\right| \mathcal{F}_n(t)dt \\ &\quad + \frac{1}{2\pi} \int_{\delta\leq|t|\leq\pi} \left|\widetilde{f}(x-t) - \widetilde{f}(x)\right| \mathcal{F}_n(t)dt. \end{aligned}$$

We estimate the two terms separately. Since $\mathcal{F}_n(t) \geq 0$ for all $t \in [-\pi, \pi]$,

$$0 \leq \frac{1}{2\pi} \int_{|t|\leq\delta} \mathcal{F}_n(t)dt \leq \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathcal{F}_n(t)dt = 1.$$

Hence,

$$\frac{1}{2\pi} \int_{|t|\leq\delta} \left|\widetilde{f}(x-t) - \widetilde{f}(x)\right| \mathcal{F}_n(t)dt \leq \frac{\varepsilon}{2} \times \frac{1}{2\pi} \int_{|t|\leq\delta} \mathcal{F}_n(t)dt \leq \frac{\varepsilon}{2}.$$

For the second term, we have

$$\frac{1}{2\pi}\int_{\delta\le|t|\le\pi}\left|\widetilde{f}(x-t)-\widetilde{f}(x)\right|\mathcal{F}_n(t)dt \le \frac{1}{2\pi}\int_{\delta\le|t|\le\pi}\frac{2M}{n\sin^2\frac{\delta}{2}}dt$$
$$\le \frac{2M}{N\sin^2\frac{\delta}{2}} < \frac{\varepsilon}{2}.$$

This shows that for all $n \ge N$,

$$|\sigma_n(x) - f(x)| < \varepsilon \qquad \text{for all } x \in [-\pi, \pi].$$

Thus, the sequence of functions $\{\sigma_n : [-\pi, \pi] \to \mathbb{C}\}$ converges uniformly to the function $f : [-\pi, \pi] \to \mathbb{C}$.

Notice that since $s_n(x)$ is in the span of $\mathcal{S}_n = \{e^{ikx} \mid -n \le n \le n\}$, $\sigma_{n+1}(x)$ is in the span of $\mathcal{S}_n(x)$. Now we apply Theorem 7.30 to give another proof of the Weierstrass approximation.

Theorem 7.31 Weierstrass Approximation Theorem

Let $f : [a, b] \to \mathbb{R}$ be a continuous function defined on $[a, b]$. Given $\varepsilon > 0$, there is a polynomial $p(x)$ such that

$$|f(x) - p(x)| < \varepsilon \qquad \text{for all } x \in [a, b].$$

Proof

It is sufficient to prove the theorem for a specific $[a, b]$. We take $[a, b] = [0, 1]$. Given $f : [0, 1] \to \mathbb{R}$ is a real-valued continuous function, we extend it to be an even function $f_e : [-1, 1] \to \mathbb{R}$, and let $g : [-\pi, \pi] \to \mathbb{R}$ be the function defined as

$$g(x) = f_e(\cos x).$$

This is well-defined since the range of $\cos x$ is $[-1, 1]$. Since $\cos x$ and $f_e : [-1, 1] \to \mathbb{R}$ are continuous even functions, $g : [-\pi, \pi] \to \mathbb{R}$ is a continuous even function. Hence, we also have $g(\pi) = g(-\pi)$.

Given $\varepsilon > 0$, Theorem 7.30 implies that there is a positive integer n such that

$$|g(x) - \sigma_{n+1}(x)| < \varepsilon. \tag{7.4}$$

Here $\sigma_{n+1}(x)$ is the $(n+1)^{\text{th}}$ Cesàro mean of the Fourier series of g. Since $g : [-\pi, \pi] \to \mathbb{R}$ is a real-valued even function, the Fourier series of g has the form

$$\frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos kx,$$

where $a_k, k \geq 0$ are real. This implies that

$$\sigma_{n+1}(x) = \sum_{k=0}^{n} \alpha_k \cos kx$$

for some real constants $\alpha_0, \alpha_1, \ldots, \alpha_n$. For any $m \geq 1$, $\cos mx$ can be written as a linear combination of $1, \cos x, \cos^2 x, \ldots, \cos^m x$. This shows that there are real constants $\beta_0, \beta_1, \ldots, \beta_n$ such that

$$\sigma_{n+1}(x) = \sum_{k=0}^{n} \beta_k \cos^k x.$$

Let

$$p(x) = \sum_{k=0}^{n} \beta_k x^k.$$

Then $\sigma_{n+1}(x) = p(\cos x)$. Thus, (7.4) says that

$$|f_e(\cos x) - p(\cos x)| < \varepsilon \qquad \text{for all } x \in [-\pi, \pi].$$

This implies that

$$|f(x) - p(x)| < \varepsilon \qquad \text{for all } x \in [0, 1],$$

which completes the proof of the theorem.

Remark 7.13

In the proof of the Weierstrass approximation theorem given above, we do not use Fourier series since the Fourier series of a 2π-periodic continuous function does not necessary converge uniformly. An example is given in [**?**]. However, there are other approaches to prove the Weierstrass approximation theorem using Fourier series. For example, one can approximate a continuous function uniformly by a continuous piecewise linear function first. The Fourier series of a continuous piecewise linear function does converge uniformly to the function itself.

In the proof given above, we used the even extension f_e of the given function f. The Fourier series of $f_e(\cos x)$ is a cosine series, so that the Cesàro mean is a polynomial in $\cos x$. One can also bypass the even extension and the composition with the cosine function, using directly uniform approximation of trigonometric functions by Taylor polynomials, as asserted by the general theory of power series.

Exercises 7.4

Question 1

Consider the function $f : [-\pi, \pi]$ defined as

$$f(x) = \begin{cases} x + \pi, & \text{if } -\pi \leq x < 0, \\ x - \pi, & \text{if } 0 \leq x \leq \pi. \end{cases}$$

The Fourier series of this function has been obtained in Exercises 7.2. Does the Fourier series converge uniformly? Justify your answer.

Question 2

Study the uniform convergence of the Fourier series of the function $f : [-\pi, \pi] \to \mathbb{R}$, $f(x) = x^2$ obtained in Exercises 7.1.

Question 3

Show that the trigonometric series

$$\sum_{k=1}^{\infty} \left(\frac{2k \cos kx + 3 \sin kx}{k^4} \right)$$

defines a continuously differentiable function $F : \mathbb{R} \to \mathbb{R}$, and find the Fourier series of the function $F' : [-\pi, \pi] \to \mathbb{R}$.

7.5 Fourier Transforms

We have seen that the Fourier series of a function $f : [-L, L] \to \mathbb{C}$ defined on $[-L, L]$ is

$$\sum_{k=-\infty}^{\infty} c_k \exp\left(\frac{i\pi kx}{L}\right),$$

where the Fourier coefficients c_k, $k \in \mathbb{Z}$ are given by

$$c_k = \frac{1}{2L}\int_{-L}^{L} f(t)\exp\left(-\frac{i\pi kt}{L}\right)dt.$$

This is also the Fourier series of the $2L$-periodic extension of the function f. Substitute the expression for c_k, we find that the Fourier series can be written as

$$\frac{1}{2\pi}\sum_{k=-\infty}^{\infty}\frac{\pi}{L}\int_{-L}^{L} f(t)\exp\left(\frac{i\pi k(x-t)}{L}\right)dt. \tag{7.5}$$

Heuristically,

$$\sum_{k=-\infty}^{\infty}\frac{\pi}{L}\exp\left(\frac{i\pi kt}{L}\right)$$

can be regarded as a Riemann sum for the function $g : \mathbb{R} \to \mathbb{C}$,

$$g(\omega) = e^{i\omega t}.$$

In the limit $L \to \infty$, one obtain heuristically the integral

$$\int_{-\infty}^{\infty} e^{i\omega t}d\omega,$$

so that (7.1) becomes

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(t)e^{i\omega(x-t)}dtd\omega.$$

This motivates us to define the Fourier transform of a function $f : \mathbb{R} \to \mathbb{C}$ as

$$\widehat{f}(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt.$$

We know that under certain conditions, the Fourier series of a function would converge to the function itself. Hence, we can also explore the conditions in which

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(t)e^{i\omega(x-t)}dtd\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty}\widehat{f}(\omega)e^{i\omega x}d\omega. \tag{7.6}$$

However, now the integrals we are working with are improper integrals. Therefore, there is another convergence issue that we need to deal with. In this section, we only give a brief discussion about Fourier transforms. An in-depth analysis would require advanced tools.

We say that a function $f : \mathbb{R} \to \mathbb{C}$ is Riemann integrable if it is Riemann integrable on any compact intervals.

Definition 7.13 L^1 and L^2 Functions

Let $f : \mathbb{R} \to \mathbb{C}$ be a Riemann integrable function. We say that f is L^1 if the improper integral

$$\int_{-\infty}^{\infty} |f(x)|dx$$

is convergent. In this case, we define the L^1-norm of f as

$$\|f\|_1 = \int_{-\infty}^{\infty} |f(x)|dx.$$

We say that f is L^2 if the improper integral

$$\int_{-\infty}^{\infty} |f(x)|^2 dx$$

is convergent. In this case, we define the L^2-norm of f as

$$\|f\|_2 = \sqrt{\int_{-\infty}^{\infty} |f(x)|^2 dx}.$$

If $f : [a, b] \to \mathbb{C}$ is any Riemann integrable function, the zero-extension of f to $\mathbb{R}$ is both a L^1 and a L^2 function. As before, the L^1 and L^2 norms are semi-norms which are positive semi-definite, where there are nonzero functions that have zero norms.

Example 7.22

Consider the function $f : \mathbb{R} \to \mathbb{R}$ defined as

$$f(x) = \frac{1}{\sqrt{x^2 + 1}}.$$

The integral

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{x^2+1}} dx$$

is not convergent, but the integral

$$\int_{-\infty}^{\infty} \frac{1}{x^2+1} dx$$

is convergent. Hence, $f : \mathbb{R} \to \mathbb{R}$ is L^2 but not L^1.

Definition 7.14 Fourier Transform

Let $f : \mathbb{R} \to \mathbb{C}$ be a Riemann integrable function. The Fourier transform of f, denoted by $\mathcal{F}[f]$ or $\widehat{f}$, is defined as

$$\mathcal{F}[f](\omega) = \widehat{f}(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t} dt,$$

for all the $\omega \in \mathbb{R}$ which this improper integral is convergent.

Example 7.23

If $f : \mathbb{R} \to \mathbb{C}$ is a L^1-function, for any $\omega \in \mathbb{R}$, the integral

$$\int_{-\infty}^{\infty} f(t)e^{-i\omega t} dt$$

converges absolutely. Hence, a L^1 function has Fourier transform $\widehat{f}$ which is defined on $\mathbb{R}$. In particular, a function that vanishes outside a bounded interval has a Fourier transform that is defined for all $\omega \in \mathbb{R}$.

Proposition 7.32

Fourier transform is a linear operation. Namely, if $f : \mathbb{R} \to \mathbb{C}$ and $g : \mathbb{R} \to \mathbb{C}$ are functions that have Fourier transforms, then for any complex numbers α and β, the function $\alpha f + \beta g : \mathbb{R} \to \mathbb{C}$ also has Fourier transform, and

$$\mathcal{F}[\alpha f + \beta g] = \alpha \mathcal{F}[f] + \beta \mathcal{F}[g].$$

Remark 7.14

In engineering, it is customary to use t as the independent variable for the function $f : \mathbb{R} \to \mathbb{C}$, and ω as the independent variable for its Fourier transform $\widehat{f} : \mathbb{R} \to \mathbb{C}$. The function f is usually a function of time t, and its Fourier transform is a function of frequency ω. Hence, the Fourier transform is a transform from the time domain to the frequency domain.

Example 7.24

Let a and b be two real numbers with $a < b$. Define the function $g : \mathbb{R} \to \mathbb{R}$ by

$$g(t) = \begin{cases} 1, & \text{if } a \leq t \leq b, \\ 0, & \text{otherwise.} \end{cases}$$

Find the Fourier transform of g.

Solution

The Fourier transform of g is

$$\widehat{g}(\omega) = \int_a^b e^{-i\omega t} dt = \begin{cases} \dfrac{i(e^{-i\omega b} - e^{-i\omega a})}{\omega}, & \text{if } \omega \neq 0, \\ b - a, & \text{if } \omega = 0. \end{cases}$$

Of special interest is when $g : \mathbb{R} \to \mathbb{R}$ is given by

$$g(t) = \begin{cases} 1, & \text{if } -a \leq t \leq a, \\ 0, & \text{otherwise,} \end{cases} \tag{7.7}$$

which is an even function. Example 7.3 shows that its Fourier transform is

$$\widehat{g}(w) = \frac{2 \sin a\omega}{\omega}.$$

One can show that this function is not L^1 but is L^2.

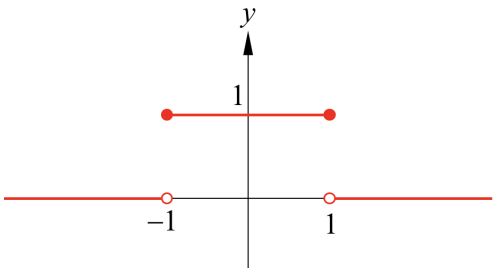


Figure 7.13: The function $g : \mathbb{R} \to \mathbb{R}$ defined by (7.3) with $a = 1$.

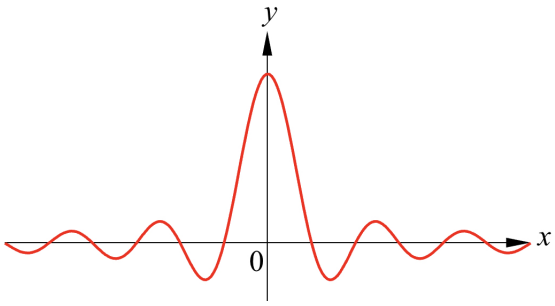


Figure 7.14: The Fourier transform of the function $g : \mathbb{R} \to \mathbb{R}$ defined by (7.3) with $a = 1$.

Remark 7.15

A function that vanishes outside a bounded interval is said to have compact support. In general, the support of a function $f : \mathbb{R} \to \mathbb{C}$ is defined to be the closure of the set of those points x such that $f(x) \neq 0$. Namely,

$$(\text{support } f) = \overline{\{x \in \mathbb{R} \,|\, f(x) \neq 0\}}.$$

Since a set is bounded if and only if its closure is bounded, a function f has compact support if and only if the set of points where f does not vanish is bounded.

Let us look at Fourier transforms of functions that does not have compact support.

Example 7.25

Let a be a positive number, and let $f : \mathbb{R} \to \mathbb{R}$ be the function defined as $f(t) = e^{-a|t|}$. Find the Fourier transform of f.

Solution

The Fourier transform of f is given by

$$\begin{aligned}
\widehat{f}(\omega) &= \int_{-\infty}^{\infty} e^{-a|t|} e^{-i\omega t} dt \\
&= \lim_{L\to\infty} \int_0^L e^{-at} \left(e^{-i\omega t} + e^{i\omega t}\right) dt \\
&= \lim_{L\to\infty} \int_0^L \left(e^{-(a+i\omega)t} + e^{-(a-i\omega)t}\right) dt \\
&= \lim_{L\to\infty} \left[-\frac{e^{-(a+i\omega)t}}{a+i\omega} - \frac{e^{-(a-i\omega)t}}{a-i\omega}\right]_0^L \\
&= \frac{1}{a+i\omega} + \frac{1}{a-i\omega} \\
&= \frac{2a}{a^2+\omega^2}.
\end{aligned}$$

Notice that the function $\widehat{f} : \mathbb{R} \to \mathbb{C}$, $\widehat{f}(\omega) = \dfrac{2a}{a^2+\omega^2}$ is a function that is both L^1 and L^2.

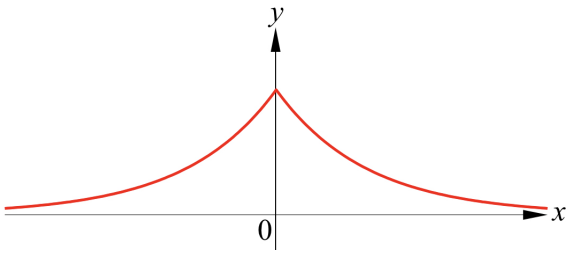


Figure 7.15: The function $f : \mathbb{R} \to \mathbb{R}$, $f(t) = e^{-|t|}$.

A function $f : \mathbb{R} \to \mathbb{R}$ of the form

$$f(x) = c \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{7.8}$$

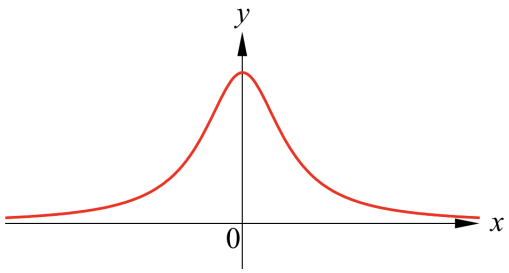


Figure 7.16: The function $\widehat{f} : \mathbb{R} \to \mathbb{R}$, $\widehat{f}(\omega) = \dfrac{2}{1+\omega^2}$, which is the Fourier transform of $f : \mathbb{R} \to \mathbb{R}$, $f(t) = e^{-|t|}$.

is the probability density function of a normal distribution with mean μ and standard deviation σ when

$$c = \frac{1}{\sqrt{2\pi}\sigma}.$$

It is also known as a *Gaussian function*. These functions are infinitely differentiable and they decay exponentially to 0 when x gets large. When $\mu = 0$ and $\sigma = 1$,

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

is the probability density of the standard normal distribution. The Fourier transform of the Gaussian function $f(t) = \exp\left(-\dfrac{t^2}{2}\right)$ is

$$\begin{aligned}
\widehat{f}(\omega) &= \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} e^{-i\omega t} dt \\
&= e^{-\frac{\omega^2}{2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(t+i\omega)^2\right) dt \\
&= e^{-\frac{\omega^2}{2}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt \\
&= \sqrt{2\pi} e^{-\frac{\omega^2}{2}}.
\end{aligned}$$

In the computation, the equality

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(t+i\omega)^2\right) dt = \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt$$

can be understood in complex analysis as shifting contours of integrations. We leave the details to the students.

Notice that for the function $f(t) = \exp\left(-\frac{t^2}{2}\right)$, its Fourier transform $\widehat{f}(\omega)$ is equal to $f(\omega)$ multiplied by $\sqrt{2\pi}$. Namely,

$$\widehat{f}(\omega) = \sqrt{2\pi} f(\omega).$$

The factor $\sqrt{2\pi}$ here is due to our normalization. Different textbooks use different conventions for Fourier transforms. Among them are the followings:

$$\begin{array}{ll} \displaystyle\int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt, & \displaystyle\int_{-\infty}^{\infty} f(t)e^{i\omega t}dt, \\ \displaystyle\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt, & \displaystyle\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(t)e^{i\omega t}dt, \\ \displaystyle\frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt, & \displaystyle\frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)e^{i\omega t}dt. \end{array}$$

Some might also replace $i\omega t$ by $2\pi i\omega t$. When one is reading about Fourier transforms, it is important to check the definition of Fourier transform that is being used.

One can show that in our definition, the Fourier transform of the Gaussian function $f(t) = e^{-at^2}$ with $a > 0$ is

$$\widehat{f}(\omega) = \int_{-\infty}^{\infty} e^{-at^2} e^{-i\omega t} dt = \sqrt{\frac{\pi}{a}} e^{-\frac{\omega^2}{4a}},$$

so that

$$\frac{1}{\sqrt{2\pi}}\widehat{f}(\omega) = \frac{1}{\sqrt{2a}} e^{-\frac{\omega^2}{4a}}.$$

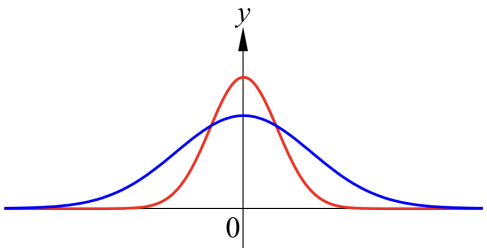


Figure 7.17: The function $f(t) = e^{-t^2}$ and the function $g(\omega) = \frac{1}{\sqrt{2\pi}}\widehat{f}(\omega) = \frac{1}{\sqrt{2}} e^{-\frac{\omega^2}{4}}$.

Remark 7.16

If $f : \mathbb{R} \to \mathbb{C}$ is a L^1 function and

$$\|f\|_1 = \int_{-\infty}^{\infty} |f(t)|dt = 0,$$

we say that f is L^1-equivalent to the zero function. If $f : \mathbb{R} \to \mathbb{C}$ is L^1-equivalent to the zero function, then for any $\omega \in \mathbb{R}$,

$$\left|\widehat{f}(\omega)\right| = \left|\int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt\right| \leq \int_{-\infty}^{\infty} |f(t)e^{-i\omega t}|dt = \int_{-\infty}^{\infty} |f(t)|dt = 0.$$

Thus, the Fourier transform of f is identically zero.

Example 7.3 shows that the Fourier transform of a L^1 function is not necessary L^1. Nevertheless, we have the following, which is an extension of the Riemann-Lebesgue lemma to L^1 functions on $\mathbb{R}$.

Theorem 7.33 Extended Riemann-Lebesgue Lemma

If the function $f : \mathbb{R} \to \mathbb{C}$ is L^1, then

$$\lim_{\beta\to\infty} \int_{-\infty}^{\infty} f(t)e^{i\beta t}dt = 0.$$

In other words, the Fourier transform $\widehat{f} : \mathbb{R} \to \mathbb{C}$ of f is a function satisfying

$$\lim_{\omega\to\pm\infty} \widehat{f}(\omega) = 0.$$

Proof

We are given that

$$J = \int_{-\infty}^{\infty} |f(t)|dt < \infty.$$

Given $\varepsilon > 0$, there is a $L > 0$ such that

$$\int_{|t|\geq L} |f(t)|dt < \frac{\varepsilon}{2}.$$

By triangle inequality, we have

$$\left|\int_{-\infty}^{\infty} f(t)e^{i\beta t}dt\right| \leq \left|\int_{-L}^{L} f(t)e^{i\beta t}dt\right| + \left|\int_{|t|\geq L} f(t)e^{i\beta t}dt\right|.$$

For the second term, we have

$$\left|\int_{|t|\geq L} f(t)e^{i\beta t}dt\right| \leq \int_{|t|\geq L} |f(t)e^{i\beta t}|dt = \int_{|t|\geq L} |f(t)|dt < \frac{\varepsilon}{2}.$$

By the Riemann-Lebesgue lemma,

$$\lim_{\beta\to 0}\int_{-L}^{L} f(t)e^{i\beta t}dt = 0.$$

Therefore, there exists $M > 0$ such that if $\beta > M$, then

$$\left|\int_{-L}^{L} f(t)e^{i\beta t}dt\right| < \frac{\varepsilon}{2}.$$

It follows that for all $\beta > M$,

$$\left|\int_{-\infty}^{\infty} f(t)e^{i\beta t}dt\right| < \varepsilon.$$

This proves the assertion.

The following theorem imposes a strong condition on a function $g : \mathbb{R} \to \mathbb{C}$ to be the Fourier transform of a L^1 function $f : \mathbb{R} \to \mathbb{C}$.

Theorem 7.34

If $f : \mathbb{R} \to \mathbb{C}$ is a L^1 function, then its Fourier transform $\widehat{f} : \mathbb{R} \to \mathbb{C}$ is uniformly continuous.

Proof

We are given that

$$J = \int_{-\infty}^{\infty} |f(t)|dt < \infty.$$

Without loss of generality, we can assume that $J > 0$. Notice that for any ω_1 and ω_2 in $\mathbb{R}$,

$$\widehat{f}(\omega_1) - \widehat{f}(\omega_2) = \int_{-\infty}^{\infty} f(t) \left(e^{-i\omega_1 t} - e^{-i\omega_2 t}\right) dt.$$

Given $\varepsilon > 0$, there is a $L > 0$ such that

$$\int_{|t|\geq L} |f(t)|dt < \frac{\varepsilon}{3}.$$

By triangle inequality, we have

$$\begin{aligned}
&\left|\widehat{f}(\omega_1) - \widehat{f}(\omega_2)\right| \leq \int_{-\infty}^{\infty} |f(t)| \left|e^{-i\omega_1 t} - e^{-i\omega_2 t}\right| dt \\
&= \int_{|t|\leq L} |f(t)| \left|e^{-i\omega_1 t} - e^{-i\omega_2 t}\right| dt + \int_{|t|\geq L} |f(t)| \left|e^{-i\omega_1 t} - e^{-i\omega_2 t}\right| dt.
\end{aligned}$$

The second term is easy to estimate since $|e^{-i\omega_1 t} - e^{-i\omega_2 t}| \leq 2$. We have

$$\int_{|t|\geq L} |f(t)| \left|e^{-i\omega_1 t} - e^{-i\omega_2 t}\right| dt \leq 2 \int_{|t|\geq L} |f(t)|dt < \frac{2\varepsilon}{3}.$$

Since the function $g : \mathbb{R} \to \mathbb{C}$, $g(u) = e^{iu}$ is continuous at $u = 0$, there exists a $\delta > 0$ such that if $|u| < \delta$, then

$$|e^{iu} - 1| < \frac{\varepsilon}{3J}.$$

Thus, given ω_1 and ω_2 in $\mathbb{R}$, if $|\omega_1 - \omega_2| < \dfrac{\delta}{L}$, then for any $t \in [-L, L]$,

$$|(\omega_1 - \omega_2)t| \leq L|\omega_1 - \omega_2| < \delta.$$

It follows that

$$|e^{-i\omega_1 t} - e^{-i\omega_2 t}| = |e^{i(\omega_1 - \omega_2)t} - 1| < \frac{\varepsilon}{3J}.$$

Therefore,

$$\int_{|t|\leq L} |f(t)| \left|e^{-i\omega_1 t} - e^{-i\omega_2 t}\right| dt \leq \frac{\varepsilon}{3J} \int_{|t|\leq L} |f(t)|dt \leq \frac{\varepsilon}{3}.$$

This proves that whenever $|\omega_1 - \omega_2| < \dfrac{\delta}{L}$, then

$$\left|\widehat{f}(\omega_1) - \widehat{f}(\omega_2)\right| < \varepsilon.$$

Hence, $\widehat{f} : \mathbb{R} \to \mathbb{C}$ is uniformly continuous.

Example 7.26

By Example 7.3, the Fourier transform of the L^1 function $f : \mathbb{R} \to \mathbb{C}$,

$$f(t) = \begin{cases} 1, & \text{if } -1 \le t \le 1, \\ 0, & \text{otherwise,} \end{cases}$$

is $\widehat{f}(\omega) = \dfrac{2\sin\omega}{\omega}$. Theorem 7.3 then implies that the function $g : \mathbb{R} \to \mathbb{R}$,

$$g(x) = \frac{\sin x}{x}$$

is uniformly continuous.

Motivated by the heuristics (7.2) from the theory of Fourier series, we make the following definition.

Definition 7.15 Inverse Fourier Transform

Given a Riemann integrable function $\widehat{f} : \mathbb{R} \to \mathbb{C}$, we define its inverse Fourier transform by

$$\mathcal{F}^{-1}[\widehat{f}](t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widehat{f}(\omega) e^{i\omega t} d\omega$$

for all the $t \in \mathbb{R}$ where this integral is convergent.

Notice that if $\widehat{f} : \mathbb{R} \to \mathbb{C}$ is a L^1 function, then $\mathcal{F}^{-1}[\widehat{f}](t)$ exists for all $t \in \mathbb{R}$, and

$$\mathcal{F}^{-1}[\widehat{f}](t) = \frac{1}{2\pi}\mathcal{F}[\widehat{f}](-t).$$

In other words, we have the following.

Proposition 7.35

Let $f : \mathbb{R} \to \mathbb{C}$ be a Riemann integrable function. If $f : \mathbb{R} \to \mathbb{C}$ has Fourier transform $\widehat{f} : \mathbb{R} \to \mathbb{C}$, and the function $\widehat{f} : \mathbb{R} \to \mathbb{C}$ has inverse Fourier transform given by $h : \mathbb{R} \to \mathbb{C}$, then the function $g : \mathbb{R} \to \mathbb{C}$, $g(t) = \widehat{f}(t)$ has a Fourier transform given by

$$\widehat{g}(\omega) = 2\pi h(-\omega).$$

Example 7.27

For the function $f : \mathbb{R} \to \mathbb{C}$, $f(t) = e^{-at^2}$, its Fourier transform is $\widehat{f} : \mathbb{R} \to \mathbb{C}$,

$$\hat{f}(\omega) = \sqrt{\frac{\pi}{a}} e^{-\frac{\omega^2}{4a}}.$$

Therefore,

$$\mathcal{F}^{-1}[\widehat{f}](t) = \frac{1}{2\sqrt{\pi a}} \int_{-\infty}^{\infty} e^{-\frac{\omega^2}{4a}} e^{i\omega t} d\omega = e^{-at^2}.$$

For $g(t) = \sqrt{\frac{\pi}{a}} e^{-\frac{t^2}{4a}}$,

$$\widehat{g}(\omega) = \sqrt{\frac{\pi}{a}} \times \sqrt{4\pi a} \exp\left(-\frac{\omega^2}{4 \times \frac{1}{4a}} \right) = 2\pi e^{-a\omega^2} = 2\pi \mathcal{F}^{-1}[\widehat{f}](-\omega).$$

Example 7.6 shows that for the function $f : \mathbb{R} \to \mathbb{C}$, $f(t) = e^{-at^2}$, we have

$$\mathcal{F}^{-1}[\widehat{f}](t) = f(t).$$

In general, we are interested in the following. If $f : \mathbb{R} \to \mathbb{C}$ is a Riemann integrable function with Fourier transform $\widehat{f} : \mathbb{R} \to \mathbb{C}$, under what conditions does $\mathcal{F}^{-1}[\widehat{f}](t)$ exists and

$$\mathcal{F}^{-1}[\widehat{f}](t) = f(t)? \tag{7.9}$$

Example 7.28

For the function $g : \mathbb{R} \to \mathbb{C}$,

$$g(\omega) = \frac{2a}{a^2 + \omega^2} \qquad \text{with } a > 0,$$

one can use contour integration techniques in complex analysis to show that when $t \in \mathbb{R}$,

$$\mathcal{F}^{-1}[g](t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{2ae^{i\omega t}}{a^2 + \omega^2} d\omega = e^{-a|t|}.$$

Hence, for the function $f : \mathbb{R} \to \mathbb{C}$, $f(t) = e^{-a|t|}$, we also have

$$\mathcal{F}^{-1}[\widehat{f}](t) = f(t) \qquad \text{for all } t \in \mathbb{R}.$$

Definition 7.16 Fourier Transform Pairs

If $f : \mathbb{R} \to \mathbb{C}$ and $g : \mathbb{R} \to \mathbb{C}$ are Riemann integrable functions, and

$$\mathcal{F}[f](\omega) = g(\omega), \qquad \mathcal{F}^{-1}[g](t) = f(t),$$

then we call the pair of functions (f, g) a Fourier transform pair.

Example 7.29

For $a > 0$, let $f : \mathbb{R} \to \mathbb{C}$ be the function $f(t) = e^{-a|t|}$, and let $g : \mathbb{R} \to \mathbb{C}$ be the function $g(t) = \dfrac{2a}{a^2 + t^2}$. Then Example 7.7 says that (f, g) is a Fourier transform pair.

The following is important for the proofs later.

Theorem 7.36

The function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \dfrac{\sin x}{x}$ is an infinitely differentiable even function that satisfies

$$\int_0^{\infty} f(x)dx = \int_0^{\infty} \frac{\sin x}{x} dx = \frac{\pi}{2}.$$

The fact that $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \dfrac{\sin x}{x}$ is infinitely differentiable has been established earlier. The formula for the improper integral can be proved using contour integration techniques and the fact that the function $g(z) = \dfrac{e^{iz}}{z}$ has a simple pole at $z = 0$ with residue 1. See for example [**?**].

Corollary 7.37

For any $a > 0$,

$$\lim_{L\to\infty} \int_0^a \frac{\sin Lx}{x} dx = \frac{\pi}{2}.$$

Proof

Making a change of variables, we have

$$\int_0^a \frac{\sin Lx}{x} dx = \int_0^{aL} \frac{\sin x}{x} dx.$$

Therefore,

$$\lim_{L\to\infty} \int_0^a \frac{\sin Lx}{x} dx = \lim_{L\to\infty} \int_0^{aL} \frac{\sin x}{x} dx = \int_0^{\infty} \frac{\sin x}{x} dx = \frac{\pi}{2}.$$

Now we can prove our main theorem. A function $f : \mathbb{R} \to \mathbb{C}$ is said to be strongly piecewise differentiable if it is strongly piecewise differentiable on any compact intervals.

Theorem 7.38 Fourier Inversion Theorem

Let $f : \mathbb{R} \to \mathbb{C}$ be a L^1-function that is strongly piecewise differentiable, and let

$$\widehat{f}(\omega) = \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt$$

be its Fourier transform. Then for any $x \in \mathbb{R}$,

$$\lim_{L\to\infty} \frac{1}{2\pi} \int_{-L}^{L} \widehat{f}(\omega) e^{i\omega x} d\omega = \frac{f_+(x) + f_-(x)}{2}.$$

Proof

Notice that

$$\begin{aligned}\int_{-L}^{L} \widehat{f}(\omega) e^{i\omega x} d\omega &= \int_{-L}^{L} \int_{-\infty}^{\infty} f(t) e^{i\omega(x-t)} dt d\omega \\ &= \int_{-L}^{L} \int_{-\infty}^{\infty} f(x-t) e^{i\omega t} dt d\omega.\end{aligned}$$

To continue, we need a technical lemma which guarantees we can interchange the order of integrations.

Lemma 7.39

Let $f : \mathbb{R} \to \mathbb{C}$ be a function that satisfies the conditions in Theorem 7.7. Then for any $L > 0$, we have

$$\int_{-L}^{L} \int_{-\infty}^{\infty} f(x-t) e^{i\omega t} dt d\omega = \int_{-\infty}^{\infty} \int_{-L}^{L} f(x-t) e^{i\omega t} d\omega dt.$$

Assuming this lemma, we can continue with the proof of Theorem 7.7.

Proof of Theorem 7.7 Continued

By Lemma 7.8, we have

$$\int_{-L}^{L} \widehat{f}(\omega) e^{i\omega x} d\omega = \int_{-\infty}^{\infty} \int_{-L}^{L} f(x-t) e^{i\omega t} d\omega dt.$$

Now we can integrate the integral with respect to ω and obtain

$$\int_{-L}^{L} \widehat{f}(\omega) e^{i\omega x} d\omega = 2 \int_{-\infty}^{\infty} f(x-t) \frac{\sin Lt}{t} dt.$$

Using the fact that $\dfrac{\sin Lt}{t}$ is an even function, we find that

$$\int_{-L}^{L} \widehat{f}(\omega) e^{i\omega x} d\omega = 2 \int_{0}^{\infty} \frac{f(x+t) + f(x-t)}{t} \sin Lt dt.$$

Split the integral into two parts, we have

$$\int_{-L}^{L} \widehat{f}(\omega) e^{i\omega x} d\omega = 2\int_0^1 \frac{f(x+t)+f(x-t)}{t} \sin Ltdt + 2\int_1^\infty \frac{f(x+t)+f(x-t)}{t} \sin Ltdt.$$

Let

$$u = \frac{f_+(x)+f_-(x)}{2}.$$

As in the proof of Lemma 7.12, the function $h : [0,1] \to \mathbb{C}$ with

$$h(t) = \frac{f(x+t)+f(x-t)-2u}{t} \qquad \text{when } t \in (0,1]$$

is a Riemann integrable function. Thererfore, the Riemann-Lebesgue lemma implies that

$$\lim_{L\to\infty} \int_0^1 h(t) \sin Ltdt = 0.$$

It follows from Corollary 7.6 that

$$\lim_{L\to\infty} 2\int_0^1 \frac{f(x+t)+f(x-t)}{t} \sin Ltdt = \lim_{L\to\infty} 4u\int_0^1 \frac{\sin Lt}{t} dt = 2\pi u.$$

On the other hand,

$$\int_1^\infty \left|\frac{f(x+t)+f(x-t)}{t}\right| dt \le 2\int_{-\infty}^\infty |f(t)| dt < \infty.$$

By the extended Riemann-Lebesgue lemma,

$$\lim_{L\to\infty} 2\int_1^\infty \frac{f(x+t)+f(x-t)}{t} \sin Ltdt = 0.$$

This completes the proof that

$$\lim_{L\to\infty} \frac{1}{2\pi}\int_{-L}^{L} \widehat{f}(\omega) e^{i\omega x} d\omega = u = \frac{f_+(x)+f_-(x)}{2}.$$

Now we prove Lemma 7.8.

Proof of Lemma 7.8

Given $\varepsilon > 0$, since

$$\int_{-\infty}^{\infty} |f(t)|dt < \infty,$$

there is an $M > 0$ such that

$$\int_{|t|\geq M} |f(t)|dt < \frac{\varepsilon}{4L}.$$

Since $e^{i\omega t}$ is an infinitely differentiable function, and $f(t)$ is a piecewice continuous function on any compact intervals, Fubini's theorem implies that

$$\int_{-L}^{L}\int_{x-M}^{x+M} f(x-t)e^{i\omega t}dtd\omega = \int_{x-M}^{x+M}\int_{-L}^{L} f(x-t)e^{i\omega t}d\omega dt.$$

Now,

$$\begin{aligned}&\left|\int_{-L}^{L}\int_{-\infty}^{\infty} f(x-t)e^{i\omega t}dtd\omega - \int_{-L}^{L}\int_{x-M}^{x+M} f(x-t)e^{i\omega t}dtd\omega\right| \\ &\leq \int_{-L}^{L}\int_{|x-t|\geq M} |f(x-t)|dtd\omega \leq 2L\int_{|t|\geq M} |f(t)|dt < \frac{\varepsilon}{2}.\end{aligned}$$

On the other hand, since $|\sin Lt| \leq L|t|$ for all $t \in \mathbb{R}$, we have

$$\begin{aligned}&\left|\int_{-\infty}^{\infty}\int_{-L}^{L} f(x-t)e^{i\omega t}d\omega dt - \int_{x-M}^{x+M}\int_{-L}^{L} f(x-t)e^{i\omega t}d\omega dt\right| \\ &\leq 2\int_{|x-t|\geq M} |f(x-t)|\left|\frac{\sin Lt}{t}\right| dt \leq 2L\int_{|t|\geq M} |f(t)|dt < \frac{\varepsilon}{2}.\end{aligned}$$

This proves that

$$\left|\int_{-L}^{L}\int_{-\infty}^{\infty} f(x-t)e^{i\omega t}dtd\omega - \int_{-\infty}^{\infty}\int_{-L}^{L} f(x-t)e^{i\omega t}d\omega dt\right| < \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, the assertion follows.

Corollary 7.40

Let $f : \mathbb{R} \to \mathbb{C}$ be a L^1 function that is continuous and strongly piecewise differentiable, and let

$$\widehat{f}(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt$$

be its Fourier transform. If $\widehat{f} : \mathbb{R} \to \mathbb{C}$ is also a L^1 function, then for any $t \in \mathbb{R}$,

$$\mathcal{F}^{-1}[\widehat{f}](t) = f(t).$$

Example 7.30

Since the function $g : \mathbb{R} \to \mathbb{R}$,

$$g(t) = \begin{cases} 1, & \text{if } -a \leq t \leq a, \\ 0, & \text{otherwise,} \end{cases}$$

is strongly piecewise differentiable L^1 function with Fourier transform

$$\widehat{g}(\omega) = \frac{2\sin a\omega}{\omega},$$

the Fourier inversion theorem implies that for $|t| < a$,

$$\lim_{L\to\infty} \frac{1}{\pi} \int_{-L}^{L} \frac{\sin a\omega}{\omega} e^{i\omega t}d\omega = 1,$$

while if $|t| > a$,

$$\lim_{L\to\infty} \frac{1}{\pi} \int_{-L}^{L} \frac{\sin a\omega}{\omega} e^{i\omega t}d\omega = 0,$$

and for $|t| = a$,

$$\lim_{L\to\infty} \frac{1}{\pi} \int_{-L}^{L} \frac{\sin a\omega}{\omega} e^{i\omega t}d\omega = \frac{1}{2}.$$

If $f : \mathbb{R} \to \mathbb{C}$ and $g : \mathbb{R} \to \mathbb{C}$ are L^2 functions, then the Cauchy Schwarz

inequality implies that for any $L > 0$,

$$\left(\int_{-L}^{L} f(t)\overline{g(t)}dt\right)^2 \leq \left(\int_{-L}^{L} |f(t)|^2 dt\right)\left(\int_{-L}^{L} |g(t)|^2 dt\right)$$
$$\leq \left(\int_{-\infty}^{\infty} |f(t)|^2 dt\right)\left(\int_{-\infty}^{\infty} |g(t)|^2 dt\right).$$

This implies that the improper integral

$$\int_{-\infty}^{\infty} f(t)\overline{g(t)}dt$$

converges absolutely. Thus, we can define a positive semi-definite inner product on the space of L^2 functions on $\mathbb{R}$ by

$$\langle f, g\rangle = \int_{-\infty}^{\infty} f(t)\overline{g(t)}dt.$$

The L^2 semi-norm is the norm induced by this inner product.

The following is a generalization of the Parseval's identity to Fourier transforms.

Theorem 7.41 Parseval- Plancherel Identity

If $f : \mathbb{R} \to \mathbb{C}$ is a Riemann integrable function that is both L^1 and L^2, then its Fourier transform $\widehat{f} : \mathbb{R} \to \mathbb{C}$ is a L^2 function. Moreover,

$$\|f\|_2^2 = \frac{1}{2\pi}\|\widehat{f}\|_2^2.$$

Namely,

$$\int_{-\infty}^{\infty} |f(t)|^2 dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} |\widehat{f}(\omega)|^2 d\omega. \tag{7.10}$$

Sketch of Proof

A rigorous proof of this theorem requires advanced tools in analysis. We give a heuristic argument for the validity of the formula (7.6) under the additional assumption that $f : \mathbb{R} \to \mathbb{C}$ is continuous and strongly piecewise differentiable, and $\widehat{f} : \mathbb{R} \to \mathbb{C}$ is also L^1.

Since $\widehat{f} : \mathbb{R} \to \mathbb{C}$ is a continuous and strongly piecewise differentiable L^1 function, the Fourier inversion theorem implies that for all $t \in \mathbb{R}$,

$$f(t) = \mathcal{F}^{-1}[\widehat{f}](t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(\omega) e^{i\omega t} dt.$$

Notice that $\widehat{f} : \mathbb{R} \to \mathbb{C}$ is a L^2-function if the limit

$$\lim_{L\to\infty} \int_{-L}^{L} \left|\widehat{f}(\omega)\right|^2 d\omega$$

exists. By the definition of Fourier transform,

$$\int_{-L}^{L} \left|\widehat{f}(\omega)\right|^2 d\omega = \int_{-L}^{L} \overline{\widehat{f}(\omega)} \int_{-\infty}^{\infty} f(t) e^{-it\omega} dt d\omega.$$

By Theorem 7.3, $\widehat{f}(\omega)$ is uniformly continuous. By the Riemann-Lebesgue lemma, $\lim\limits_{\omega\to\pm\infty} \widehat{f}(\omega) = 0$. These imply that the function $\widehat{f} : \mathbb{R} \to \mathbb{C}$ is bounded. Using the same reasoning as in the proof of Lemma 7.8, we can interchange the order of integrations and obtain

$$\int_{-L}^{L} \left|\widehat{f}(\omega)\right|^2 d\omega = \int_{-\infty}^{\infty} f(t) \int_{-L}^{L} \overline{\widehat{f}(\omega)} e^{-it\omega} d\omega dt.$$

Since $\widehat{f} : \mathbb{R} \to \mathbb{C}$ is L^1, we can take the $L \to \infty$ limit under the integral sign. Since

$$\lim_{L\to\infty} \int_{-L}^{L} \overline{\widehat{f}(\omega)} e^{-it\omega} d\omega = \overline{\int_{-\infty}^{\infty} \widehat{f}(\omega) e^{i\omega t} d\omega},$$

we conclude that

$$\lim_{L\to\infty} \int_{-L}^{L} \left|\widehat{f}(\omega)\right|^2 d\omega = 2\pi \int_{-\infty}^{\infty} |f(t)|^2 dt.$$

Example 7.31

For the function $f : \mathbb{R} \to \mathbb{C}$, $f(t) = e^{-a|t|}$ with $a > 0$, its Fourier transform is $\widehat{f} : \mathbb{R} \to \mathbb{C}$, $\widehat{f}(\omega) = \dfrac{2a}{a^2 + \omega^2}$. Notice that

$$\int_{-\infty}^{\infty} |f(t)|^2 dt = 2\int_0^{\infty} e^{-2at} dt = \frac{1}{a}.$$

The Parseval-Plancherel formula implies that

$$\int_{-\infty}^{\infty} \frac{1}{(a^2+\omega^2)^2} d\omega = \frac{1}{4a^2}\int_{-\infty}^{\infty} \left|\widehat{f}(\omega)\right|^2 d\omega = \frac{\pi}{2a^2}\int_{-\infty}^{\infty} |f(t)|^2 dt = \frac{\pi}{2a^3}.$$

One of the applications of Fourier transforms is to solve differential equations. For this we need the following.

Theorem 7.42

Let $f : \mathbb{R} \to \mathbb{C}$ be a continuously differentiable L^1 function such that

$$\lim_{t \to \pm\infty} f(t) = 0,$$

and its derivative $f' : \mathbb{R} \to \mathbb{C}$ is a Riemann integrable function that has Fourier transform. Then

$$\mathcal{F}[f'](\omega) = i\omega \mathcal{F}[f](\omega).$$

Proof

This follows from integration by parts. For any a and b with $a < b$,

$$\int_a^b f'(t)e^{-it\omega} dt = \left[f(t)e^{-it\omega}\right]_a^b + i\omega \int_a^b f(t)e^{-it\omega} dt.$$

The assertion follows by taking the limit $a \to -\infty$ and $b \to \infty$.

Example 7.32

Find the Fourier transform of the function $f : \mathbb{R} \to \mathbb{C}$, $f(t) = t^2 e^{-t^2}$.

Solution

Let $g : \mathbb{R} \to \mathbb{C}$ be the function

$$g(t) = e^{-t^2}.$$

Then

$$g'(t) = -2te^{-t^2}, \qquad g''(t) = -2e^{-t^2} + 4t^2e^{-t^2} = -2g(t) + 4f(t).$$

Since

$$\lim_{t\to\pm\infty} g(t) = 0, \qquad \lim_{t\to\pm\infty} g'(t) = 0,$$

and

$$\mathcal{F}[g](\omega) = \sqrt{\pi}e^{-\frac{\omega^2}{4}},$$

Theorem 7.11 implies that

$$\mathcal{F}[g''](\omega) = -\sqrt{\pi}\omega^2 e^{-\frac{\omega^2}{4}}.$$

By linearity,

$$\mathcal{F}[g''] = -2\mathcal{F}[g] + 4\mathcal{F}[f].$$

Therefore,

$$\mathcal{F}[f](\omega) = \sqrt{\pi}\left(\frac{1}{2} - \frac{\omega^2}{4}\right)e^{-\frac{\omega^2}{4}}.$$

In the following, we consider an operation called convolution.

Definition 7.17 Convolution

Let $f : \mathbb{R} \to \mathbb{C}$ and $g : \mathbb{R} \to \mathbb{C}$ be Riemann integrable functions. The convolution of f and g is the function defined as

$$(f * g)(x) = \int_{-\infty}^{\infty} f(x-t)g(t)dt = \int_{-\infty}^{\infty} f(t)g(x-t)dt,$$

whenever this integral is convergent.

Notice that the improper integral defining $f * g$ is convergent for any x in $\mathbb{R}$ when f and g are L^2 functions. Convolutions can be defined for a wider class of functions. For example, if the supports of the functions f and g are both contained

in $[0, \infty)$, then the integral is only nonzero when $0 \leq t \leq x$. This gives

$$(f * g)(x) = \int_0^x f(t)g(x-t)dt,$$

which is also well-defined for any $x \in \mathbb{R}$. In fact, this is the convolution one sees in the theory of Laplace transforms.

Example 7.33

Let $f : \mathbb{R} \to \mathbb{R}$ be the function defined as

$$f(x) = \begin{cases} 1, & \text{if } 0 \leq x \leq 1, \\ 0, & \text{otherwise}, \end{cases} \tag{7.11}$$

and let $g : \mathbb{R} \to \mathbb{R}$ be the function $g(x) = x$. Then

$$(f * g)(x) = \int_{-\infty}^{\infty} f(t)g(x-t)dt = \int_0^1 (x-t)dt = x - \frac{1}{2}.$$

For $f * f$, we have

$$(f * f)(x) = \int_0^1 f(x-t)dt = \int_{x-1}^x f(t)dt = \begin{cases} 0, & \text{if } \quad x < 0, \\ x, & \text{if } 0 \leq x < 1, \\ 2-x, & \text{if } 1 \leq x \leq 2, \\ 0, & \text{if } \quad x > 2. \end{cases}$$



Figure 7.18: The function $f : \mathbb{R} \to \mathbb{R}$ defined by (7.7) and the function $f * f$.

Convolution usually smooths up a function, as shown in Figure 7.6.

In the theory of Fourier transforms, convolution plays an important role because of the following.

Theorem 7.43

Let $f : \mathbb{R} \to \mathbb{C}$ and $g : \mathbb{R} \to \mathbb{C}$ be functions that are both L^1 and L^2. Then $(f * g) : \mathbb{R} \to \mathbb{C}$ is an L^1 function and

$$\mathcal{F}[f * g] = \mathcal{F}[f]\mathcal{F}[g].$$

Sketch of Proof

Fubini's theorem implies that

$$\begin{aligned}\int_{-\infty}^{\infty} |(f * g)(x)|dx &\leq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |f(x-t)||g(t)|dtdx \\ &\leq \int_{-\infty}^{\infty} |g(t)| \int_{-\infty}^{\infty} |f(x-t)|dxdt \\ &= \|f\|_1 \int_{-\infty}^{\infty} |g(t)|dt = \|f\|_1\|g\|_1.\end{aligned}$$

This shows that $f * g$ is an L^1 function. By Fubini's theorem again, we have

$$\begin{aligned}\mathcal{F}[f * g](\omega) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x-t)g(t)dte^{-i\omega x}dx \\ &= \int_{-\infty}^{\infty} g(t) \int_{-\infty}^{\infty} f(x-t)e^{-i\omega(x-t)}dxe^{-i\omega t}dt \\ &= \mathcal{F}[f](\omega) \int_{-\infty}^{\infty} g(t)e^{-i\omega t}dt = \mathcal{F}[f](\omega)\mathcal{F}[g](\omega).\end{aligned}$$

Example 7.34

Find the Fourier transform of the function $g : \mathbb{R} \to \mathbb{R}$,

$$g(t) = \begin{cases} 0, & \text{if} \quad x < 0, \\ t, & \text{if } 0 \leq t < 1, \\ 2 - t, & \text{if } 1 \leq t \leq 2, \\ 0, & \text{if} \quad t > 2. \end{cases}$$

Solution

By Example 7.12, $g = f * f$, where f is the function given by 7.7. The Fourier transform of f is

$$\widehat{f}(\omega) = \frac{1 - e^{-i\omega}}{i\omega} = e^{-\frac{i\omega}{2}} \frac{2 \sin \frac{\omega}{2}}{\omega}.$$

Therefore, the Fourier transform of g is

$$\widehat{g}(\omega) = \widehat{f}(\omega) \times \widehat{f}(\omega) = e^{-i\omega} \frac{4 \sin^2 \frac{\omega}{2}}{\omega^2}.$$

Now we list down some other useful properties of Fourier transforms. The proofs are left as exercises.

Theorem 7.44

Let $f : \mathbb{R} \to \mathbb{C}$ be a L^1 function and let a be a real number.

(a) If $g : \mathbb{R} \to \mathbb{C}$ is the function $g(t) = f(t - a)$, then $\widehat{g}(\omega) = e^{-ia\omega} \widehat{f}(\omega)$.

(b) If $h : \mathbb{R} \to \mathbb{C}$ is the function $h(t) = f(t)e^{iat}$, then $\widehat{h}(\omega) = \widehat{f}(\omega - a)$.

Example 7.35

Let us consider solving a partial differential equation of the form

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = 0, \tag{7.12}$$

where c is a positive constant. This is the called the wave equation. The function u is a function in $(t, x) \in \mathbb{R}^2$. For simplicity, we assume that u, u_t, u_{tt} are infinitely differentiable bounded L^1 functions which decays to 0 when $t \to \pm\infty$.

Let $\widehat{u}(\omega, x)$ be the Fourier transform of u with respect to the variable t. Then

$$\mathcal{F}[u_{tt}](\omega, x) = -\omega^2 \widehat{u}(\omega, x).$$

It can be justified that

$$\mathcal{F}[u_{xx}] = \frac{\partial^2}{\partial x^2}\mathcal{F}[u].$$

Thus, under Fourier transform with respect to t, the partial differential equation (7.8) is transformed to a second order ordinary differential equation

$$\widehat{u}_{xx}(\omega, x) + \frac{\omega^2}{c^2}\widehat{u}(\omega, x) = 0 \tag{7.13}$$

with respect to the variable x. The general solution is

$$\widehat{u}(\omega, x) = A(\omega)e^{\frac{i\omega}{c}x} + B(\omega)e^{-\frac{i\omega}{c}x}$$

for some infinitely differentiable functions $A(\omega)$ and $B(\omega)$. Assume that

$$\mathcal{F}^{-1}[A](t) = \widetilde{A}(t), \qquad \mathcal{F}^{-1}[B](t) = \widetilde{B}(t).$$

Then $A(\omega)e^{\frac{i\omega}{c}x}$ and $B(\omega)e^{-\frac{i\omega}{c}x}$ are the Fourier transforms of the functions

$$\widetilde{A}\left(t + \frac{x}{c}\right) \quad \text{and} \quad \widetilde{B}\left(t - \frac{x}{c}\right)$$

respectively. These give

$$u(t, x) = \widetilde{A}\left(t + \frac{x}{c}\right) + \widetilde{B}\left(t - \frac{x}{c}\right).$$

Let

$$\phi(t) = \widetilde{A}\left(\frac{t}{c}\right) \quad \text{and} \quad \psi(t) = \widetilde{B}\left(-\frac{t}{c}\right).$$

Then

$$u(t, x) = \phi(x + ct) + \psi(x - ct).$$

This shows that the solution of the wave equation can be written as a sum of a left-travelling wave $\phi(x + ct)$ and a right-travelling wave $\psi(x - ct)$.

Exercises 7.5

Question 1

Let $f : \mathbb{R} \to \mathbb{C}$ be a L^1 function and let a be a real number. Define the function $g : \mathbb{R} \to \mathbb{C}$ by

$$g(t) = f(t - a).$$

Show that

$$\widehat{g}(\omega) = e^{-ia\omega}\widehat{f}(\omega).$$

Question 2

Let $f : \mathbb{R} \to \mathbb{C}$ be a L^1 function and let a be a real number. Define the function $g : \mathbb{R} \to \mathbb{C}$ by

$$g(t) = f(t)e^{iat}.$$

Show that

$$\widehat{g}(\omega) = \widehat{f}(\omega - a).$$

Question 3

Find the Fourier transform of the function $f : \mathbb{R} \to \mathbb{C}$.

(a) $f(t) = \dfrac{1}{t^2 + 4}$

(b) $f(t) = \dfrac{1}{t^2 + 4t + 13}$

(c) $f(t) = \dfrac{\sin t}{t^2 + 4t + 13}$

Question 4

Let $f : \mathbb{R} \to \mathbb{R}$ be the function $f(t) = e^{-3|t|}$, and let $g : \mathbb{R} \to \mathbb{R}$ be the function $g(t) = (f * f)(t)$. Use convolution theorem to find the Fourier transform of the function $g : \mathbb{R} \to \mathbb{R}$.

Question 5

Let a and b two distinct positive numbers, and let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ be the functions $f(t) = e^{-at^2}$ and $g(t) = e^{-bt^2}$. Find the function $h : \mathbb{R} \to \mathbb{R}$ defined as $h(t) = (f * g)(t)$.

Question 6

Let $f : \mathbb{R} \to \mathbb{C}$ be a bounded L^1 function. Show that f is L^2.

Appendix A

Sylvester's Criterion

In this section, we give a proof of the Sylvester's criterion, which gives a necessary and sufficient condition for a symmetric matrix to be positive definite. The proof uses the LDU factorization of a matrix.

Given an $n \times n$ matrix A and an integer $1 \leq k \leq n$, the k^{th} principal submatrix of A, denoted by $M_k(A)$, is the $k \times k$ matrix consists of the first k rows and first k columns of A. The Sylvester's criterion is the following.

Theorem A.1 Sylvester's Criterion for Positive Definiteness

An $n \times n$ symmetric matrix A is positive definite if and only if $\det M_k > 0$ for all $1 \leq k \leq n$, where M_k is its k^{th} principal submatrix.

For a positive integer n, let $\mathcal{M}_n$ be the vector space of $n \times n$ matrices, and let $\mathcal{L}_n$, $\mathcal{U}_n$ and $\mathcal{D}_n$ be respectively the subspaces that consist of lower triangular, upper triangular, and diagonal matrices. Also, let

$$\widetilde{\mathcal{L}}_n = \{L \in \mathcal{L}_n \,|\, \text{all the diagonal entries of } L \text{ are equal to } 1\},$$
$$\widetilde{\mathcal{U}}_n = \{U \in \mathcal{U}_n \,|\, \text{all the diagonal entries of } U \text{ are equal to } 1\}.$$

Notice that L is in $\widetilde{\mathcal{L}}_n$ if and only if its transpose L^T is in $\widetilde{\mathcal{U}}_n$.

The set of $n \times n$ invertible matrices is a group under matrix multiplication. This group is denoted by $\mathrm{GL}(n, \mathbb{R})$, and is called the general linear group. As a set, it is the subset of $\mathcal{M}_n$ that consists of all the matrices A with $\det A \neq 0$. The group $\mathrm{GL}(n, \mathbb{R})$ has a subgroup that contains all the invertible matrices with determinant 1, deonoted by $\mathrm{SL}(n, \mathbb{R})$, and is called the special linear group. The sets $\widetilde{\mathcal{L}}_n$ and $\widetilde{\mathcal{U}}_n$ are subgroups of $\mathrm{SL}(n, \mathbb{R})$.

If A is an $n \times n$ matrix, an LDU factorization of A is a factorization of the form

$$A = LDU,$$

where $L \in \widetilde{\mathcal{L}}_n$, $D \in \mathcal{D}_n$, and $U \in \widetilde{\mathcal{U}}_n$. Notice that $\det A = \det D$. Hence, A is invertible if and only if all the diagonal entries of D are nonzero.

The following proposition says that the LDU decomposition of an invertible matrix is unique.

Proposition A.2 Uniqueness of *LDU* Factorization

If A is an $n \times n$ invertible matrix that has an LDU factorization, then the factorization is unique.

Proof

We need to prove that if L_1, L_2 are in $\widetilde{\mathcal{L}}_n$, U_1, U_2 are in $\widetilde{\mathcal{U}}_n$, D_1, D_2 are in $\mathcal{D}_n$, and

$$L_1 D_1 U_1 = L_2 D_2 U_2,$$

then $L_1 = L_2$, $U_1 = U_2$ and $D_1 = D_2$.
Let $L = L_2^{-1} L_1$ and $U = U_2 U_1^{-1}$. Then

$$L D_1 = D_2 U.$$

Notice that L is in $\widetilde{\mathcal{L}}_n$ and LD_1 is in $\mathcal{L}_n$. Similarly, U is in $\widetilde{\mathcal{U}}_n$ and $D_2 U$ is in $\mathcal{U}_n$. The intersection of $\mathcal{L}_n$ and $\mathcal{U}_n$ is $\mathcal{D}_n$. Thus, there exists $D \in \mathcal{D}_n$ such that

$$L D_1 = D_2 U = D.$$

Since A is invertible, D_1 and D_2 are invertible. Hence,

$$L = D D_1^{-1} \quad \text{and} \quad U = D_2^{-1} D$$

are diagonal matrices. Since all the diagonal entries of L and U are 1, we find that $D D_1^{-1} = I_n$ and $D_2^{-1} D = I_n$, where I_n is the $n \times n$ identity matrix. This proves that

$$D_1 = D = D_2.$$

But then $L = I_n = U$, which imply that $L_1 = L_2$ and $U_1 = U_2$.

Corollary A.3

(i) Given $L_0 \in \mathcal{L}_n$, if L_0 is invertible, it has a unique LDU decomposition with $U = I_n$ the $n \times n$ identity matrix.

(ii) Given $U_0 \in \mathcal{U}_n$, if U_0 is invertible, it has a unique LDU decomposition with $L = I_n$ the $n \times n$ identity matrix.

Proof

It suffices to establish (i). The uniqueness is asserted in Proposition A.2. For the existence, let $L_0 = [a_{ij}]$, where $a_{ij} = 0$ if $i < j$. Since L_0 is invertible, $a_{ii} \neq 0$ for all $1 \leq i \leq n$. Let $D = [d_{ij}]$ be the diagonal matrix with $d_{ii} = a_{ii}$ for $1 \leq i \leq n$. Then D is invertible. Define $L = L_0 D^{-1}$. Then L is a lower triangular matrix and for $1 \leq i \leq n$,

$$L_{ii} = a_{ii} d_{ii}^{-1} = 1.$$

This shows that L is in $\widetilde{\mathcal{L}}_n$. Thus, $L_0 = LD$ is the LDU decomposition of L_0 with $U = I_n$.

The following lemma says that multiplying by a matrix L in $\widetilde{\mathcal{L}}_n$ does not affect the determinants of the principal submatrices.

Lemma A.4

Let A be an $n \times n$ matrix, and let L be a matrix in $\widetilde{\mathcal{L}}_n$. If $B = LA$, then for $1 \leq k \leq n$,

$$\det M_k(B) = \det M_k(A).$$

Proof

For an $n \times n$ matrix C, we partition it into four blocks

$$C = \left[\begin{array}{c|c} M_k(C) & N_k(C) \\ \hline P_k(C) & Q_k(C) \end{array}\right].$$

For $L \in \widetilde{\mathcal{L}}_n$, $M_k(L)$ is in $\widetilde{\mathcal{L}}_k$, and $N_k(L)$ is the zero matrix. Now $B = LA$ implies that

$$\left[\begin{array}{c|c} M_k(B) & N_k(B) \\ \hline P_k(B) & Q_k(B) \end{array}\right] = \left[\begin{array}{c|c} M_k(L) & N_k(L) \\ \hline P_k(L) & Q_k(L) \end{array}\right] \left[\begin{array}{c|c} M_k(A) & N_k(A) \\ \hline P_k(A) & Q_k(A) \end{array}\right].$$

This implies that

$$M_k(B) = M_k(L)M_k(A) + N_k(L)P_k(A) = M_k(L)M_k(A).$$

Since $M_k(L) \in \widetilde{\mathcal{L}}_k$, $\det M_k(L) = 1$. Therefore,

$$\det M_k(B) = \det M_k(A).$$

Lemma A.4 has an upper triangular counterpart.

Corollary A.5

Let A be an $n \times n$ matrix, and let U be a matrix in $\widetilde{\mathcal{U}}_n$. If $B = AU$, then for $1 \leq k \leq n$,

$$\det M_k(B) = \det M_k(A).$$

Sketch of Proof

Notice that that $M_k(B^T) = M_k(B)^T$, and $B^T = U^T A^T$, where U^T is in $\widetilde{\mathcal{L}}_n$. The result follows from the fact that $\det C^T = \det C$ for any $k \times k$ matrix C.

Now we prove the following theorem which asserts the existence of LDU decomposition for a matrix A with $\det M_k(A) \neq 0$ for all $1 \leq k \leq n$.

Theorem A.6

Let $A = [a_{ij}]$ be an $n \times n$ matrix such that $\det M_k(A) \neq 0$ for all $1 \leq k \leq n$. Then A has a unique LDU decomposition.

Proof

Notice that $M_n(A) = A$. Since we assume that $\det M_n(A) \neq 0$, A is invertible. The uniqueness of the LDU decomposition of A is asserted in Proposition A.2.

We prove the statement by induction on n. When $n = 1$, take $L = U = [1]$ and $D = A = [a]$ itself. Then $A = LDU$ is the LDU decomposition of A.

Let $n \geq 2$. Suppose we have proved that any $(n-1) \times (n-1)$ matrix B that satisfies $\det M_k(B) \neq 0$ for $1 \leq k \leq n-1$ has a unique LDU decomposition.

Now assume that A is an $n \times n$ matrix with $\det M_k(A) \neq 0$ for all $1 \leq k \leq n$. Since $\det M_1(A) = a_{11}$, $a = a_{11} \neq 0$. Let $L_1 = [L_{ij}]$ be the matrix in $\widetilde{\mathcal{L}}_n$ such that for $2 \leq i \leq n$,

$$L_{i1} = \frac{a_{i1}}{a},$$

and for $2 \leq j < i \leq n$, $L_{ij} = 0$. Namely,

$$L_1 = \left[\begin{array}{c|c} 1 & 0 \\ \hline P_1(L_1) & I_{n-1} \end{array}\right],$$

where

$$P_1(L_1) = \frac{1}{a} P_1(A).$$

Notice that

$$L_1^{-1} = \left[\begin{array}{c|c} 1 & 0 \\ \hline -P_1(L_1) & I_{n-1} \end{array}\right],$$

and

$$C = L_1^{-1} A = \left[\begin{array}{c|c} 1 & 0 \\ \hline -P_1(L_1) & I_{n-1} \end{array}\right] \left[\begin{array}{c|c} a & N_1(A) \\ \hline P_1(A) & Q_1(A) \end{array}\right] = \left[\begin{array}{c|c} a & N_1(C) \\ \hline P_1(C) & Q_1(C) \end{array}\right]$$

is a matrix with

$$P_1(C) = -aP_1(L_1) + P_1(A) = 0.$$

By Lemma A.4,

$$\det M_k(C) = \det M_k(A) \qquad \text{for all } 1 \le k \le n.$$

Let $B = Q_1(C)$. Then B is an $(n-1) \times (n-1)$ matrix. Since $P_1(C) = 0$, we find that for $1 \le k \le n-1$,

$$\det M_{k+1}(C) = a \det M_k(B).$$

This shows that

$$\det M_k(B) \neq 0 \qquad \text{for all } 1 \le k \le n-1.$$

By inductive hypothesis, B has a unique LDU decomposition given by

$$B = L_B D_B U_B.$$

Now let L_2 be the matrix in $\widetilde{\mathcal{L}}_n$ given by

$$L_2 = \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & L_B^{-1} \end{array}\right].$$

One can check that

$$(L_1 L_2)^{-1} A = L_2^{-1} C = \left[\begin{array}{c|c} a & N_1(C) \\ \hline 0 & D_B U_B \end{array}\right].$$

Let $L = L_1 L_2$. Then L is in $\widetilde{\mathcal{L}}_n$. Since $D_B U_B$ is an upper triangular $(n-1) \times (n-1)$ matrix, $L^{-1}A$ is an upper triangular $n \times n$ matrix. By Corollary A.3, $L^{-1}A$ has a decomposition

$$L^{-1} A = DU,$$

where $D \in \mathcal{D}_n$ and $U \in \widetilde{\mathcal{U}}_n$. Thus, $A = LDU$ is the LDU decomposition of A.

Now we can complete the proof of the Sylvester's criterion for a symmetric matrix to be positive definite.

Proof of Sylvester's Criterion

Let A be an $n \times n$ symmetric matrix. First we prove that if A is positive definite, then for $1 \leq k \leq n$, $\det M_k(A) > 0$. Notice that $M_k(A)$ is also a symmetric matrix. For $\mathbf{u} \in \mathbb{R}^k$, let $\mathbf{v}$ be the vector in $\mathbb{R}^n$ given by $\mathbf{v} = (\mathbf{u}, 0, \ldots, 0)$. Then

$$\mathbf{v}^T A \mathbf{v} = \mathbf{u}^T M_k(A) \mathbf{u}.$$

This shows that $M_k(A)$ is also positive definite. Hence, all the eigenvalues of $M_k(A)$ must be positive. This implies that $\det M_k(A) > 0$.
Conversely, assume that $\det M_k(A) > 0$ for all $1 \leq k \leq n$. By Theorem A.6, A has a LDU decomposition given by

$$A = LDU.$$

Since A is symmetric, $A^T = A$. This gives

$$U^T D^T L^T = A^T = A = LDU.$$

Since U^T is in $\widetilde{\mathcal{L}}_n$ and L^T is in $\widetilde{\mathcal{U}}_n$, the uniqueness of LDU decomposition implies that $U = L^T$. Hence,

$$A = LDL^T.$$

By Lemma A.4 and Corollary A.5,

$$\det M_k(A) = \det M_k(D).$$

If $D = [d_{ij}]$, let $\tau_i = d_{ii}$. Then $M_k(D) = \tau_1 \tau_2 \ldots \tau_k$. Since $\det M_k(A) > 0$ for all $1 \leq k \leq n$, $\tau_i > 0$ for all $1 \leq i \leq n$. By the invertible change of coordinates $\mathbf{y} = L^T \mathbf{x}$, we find that if $\mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$,

$$\mathbf{x}^T A \mathbf{x} = \mathbf{y}^T D \mathbf{y} = \tau_1 y_1^2 + \tau_2 y_2^2 + \cdots + \tau_n y_n^2 > 0.$$

This proves that A is positive definite.

Appendix B

Volumes of Parallelepipeds

In this appendix, we give a geometric proof of the formula for the volume of a parallelepiped in $\mathbb{R}^n$.

Theorem B.1

Let $\mathscr{P}$ be a parallelepiped in $\mathbb{R}^n$ spanned by the linearly independent vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$. Then the volume of $\mathscr{P}$ is equal to $|\det A|$, where A is the matrix whose column vectors are $\mathbf{v}_1, \ldots, \mathbf{v}_n$.

Let us look at a special case of parallelepiped where this theorem is easy to prove by simple geometric consideration.

Definition B.1 Generalized Rectangles

A parallelepiped that is spanned by n nonzero orthogonal vectors $\mathbf{w}_1$, …, $\mathbf{w}_n$ is called a generalized rectangle.

A generalized rectangle R based at the origin and spanned by the n nonzero orthogonal vectors $\mathbf{w}_1$, …, $\mathbf{w}_n$ is equal to $B(\mathbb{Q}_n)$, where $Q_n = [0,1]^n$ is the standard unit cube, and B is the matrix

$$B = \left[\mathbf{w}_1 \;\middle|\; \cdots \;\middle|\; \mathbf{w}_n\right].$$

By geometric consideration, the volume of R is given by the product of the lengths of its edges. Namely,

$$\text{vol}\,(R) = \|\mathbf{w}_1\| \cdots \|\mathbf{w}_n\|.$$

To see that this is equal to $\det B$, let $\mathbf{u}_1$, …, $\mathbf{u}_n$ be the unit vectors in the directions of $\mathbf{w}_1$, …, $\mathbf{w}_n$. Namely,

$$\mathbf{u}_i = \frac{\mathbf{w}_i}{\|\mathbf{w}_i\|} \qquad 1 \le i \le n.$$

Then $B = PD$, where P is an orthogonal matrix and D is a diagonal matrix given respectively by

$$P = \left[\mathbf{u}_1 \,\middle|\, \cdots \,\middle|\, \mathbf{u}_n\right], \qquad D = \begin{bmatrix} \|\mathbf{w}_1\| & 0 & \cdots & 0 \\ 0 & \|\mathbf{w}_2\| & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \|\mathbf{w}_n\| \end{bmatrix}. \tag{B.1}$$

An $n \times n$ matrix P is called an *orthogonal matrix* if

$$P^T P = P P^T = I_n,$$

where I_n is the $n \times n$ identity matrix. A matrix P is orthogonal if and only if the column vectors of P form an orthonormal basis of $\mathbb{R}^n$. If P is orthogonal, $P^{-1} = P^T$, and P^{-1} is also orthogonal. From $P^T P = I_n$, we find that

$$\det(P)\det(P^T) = \det(I_n) = 1.$$

Since $\det(P^T) = \det(P)$, we have $\det(P)^2 = 1$. Hence, the determinant of an orthogonal matrix can only be 1 or -1. Therefore, when $B = PD$, with P and D as given in (B.1), we have

$$|\det B| = |\det P \det D| = |\det D| = \|\mathbf{w}_1\| \cdots \|\mathbf{w}_n\|.$$

Remark B.1

In the argument above, we do not show that the volume of a generalized rectangle spanned by the n nonzero orthogonal vectors $\mathbf{w}_1, \ldots, \mathbf{w}_n$ is equal to $\|\mathbf{w}_1\| \cdots \|\mathbf{w}_n\|$ using the definition of $\mathrm{vol}\,(R)$ in terms of a Riemann integral $\displaystyle\int_R d\mathbf{x}$. This is elementary but tedious.

A linear transformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{T}(\mathbf{x}) = P\mathbf{x}$ defined by an orthogonal matrix P is called an *orthogonal transformation*. The significance of an orthognal transformation is as follows. For any $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$,

$$\langle \mathbf{T}(\mathbf{u}), \mathbf{T}(\mathbf{v}) \rangle = (P\mathbf{u})^T (P\mathbf{v}) = \mathbf{u}^T P^T P \mathbf{v} = \mathbf{u}^T \mathbf{v} = \langle \mathbf{u}, \mathbf{v} \rangle.$$

Namely, $\mathbf{T}$ preserves inner products. Since lengths and angles are defined in terms of the inner product, this implies that an orthogonal transformation preserves lengths and angles.

Under an orthogonal transformation, the image of a rectangle R is a rectangle that is congruent to R. Since the volume of a Jordan measurable set $\mathfrak{D}$ is obtained by taking the limit of a sequence of Darboux lower sums, and each Darboux lower sum is a sum of volumes of rectangles with disjoint interiors that lie in $\mathfrak{D}$, we find that orthogonal transformations also preserve the volumes of Jordan measurable sets.

Theorem B.2

If $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{T}(\mathbf{x}) = P\mathbf{x}$ is an orthogonal transformation, and $\mathfrak{D}$ is a Jordan measurable set, then $\mathbf{T}(\mathfrak{D})$ is also Jordan measurable and

$$\operatorname{vol}(\mathbf{T}(\mathfrak{D})) = \operatorname{vol}(\mathfrak{D}).$$

To finish the proof of Theorem B.1, we also need the following fact.

Proposition B.3

Let $\mathscr{P}$ be a parallelepiped based at the origin and spanned by the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$. Assume that

$$\pi_n(\mathbf{v}_i) = 0 \qquad \text{for } 1 \leq i \leq n-1,$$

or equivalently, $\mathbf{v}_1, \ldots, \mathbf{v}_{n-1}$ lies in the plane $x_n = 0$. For $1 \leq i \leq n-1$, let $\mathbf{z}_i \in \mathbb{R}^{n-1}$ be such that $\mathbf{v}_i = (\mathbf{z}_i, 0)$. If $\mathcal{Q}$ is the parallelepiped in $\mathbb{R}^{n-1}$ based at the origin and spanned by $\mathbf{z}_1, \ldots, \mathbf{z}_{n-1}$, then

$$\operatorname{vol}(\mathscr{P}) = \operatorname{vol}(\mathcal{Q})h,$$

where h is the distance from $\mathbf{v}_n$ to the $x_n = 0$ plane, which is given explicitly by

$$h = \left| \operatorname{proj}_{\mathbf{e}_n} \mathbf{v}_n \right|.$$

When $n = 3$, this can be argued geometrically. For general n, let us give a proof using the definition of volume as a Riemann integral.

Proof

Recall that

$$\mathscr{P} = \{t_1\mathbf{v}_1 + \cdots + t_{n-1}\mathbf{v}_{n-1} + t_n\mathbf{v}_n \,|\, \mathbf{t} \in [0,1]^n\}.$$

Notice that $\mathbf{v}_n$ can be written as $\mathbf{v}_n = (\mathbf{a}, h)$ for some $\mathbf{a} \in \mathbb{R}^{n-1}$. Hence, if a point $\mathbf{x}$ is in $\mathscr{P}$, then

$$\mathbf{x} = \left(\frac{t}{h}\mathbf{a} + \mathbf{z}, t\right),$$

where $0 \leq t \leq h$, and $\mathbf{z}$ is a point in $\mathcal{Q}$. For $0 \leq t \leq h$, let

$$\mathcal{Q}_t = \left\{ \left(\frac{t}{h}\mathbf{a} + \mathbf{z}, t\right) \,\middle|\, 0 \leq t \leq h \right\}.$$

Then it is a $(n-1)$-dimensional parallelepiped contained in the hyperplane $x_n = t$, which is a translate of the $(n-1)$-dimensional parallelepiped $\mathcal{Q}_0$. By Fubini's theorem,

$$\begin{aligned}\text{vol}\,(\mathscr{P}) &= \int_{\mathscr{P}} d\mathbf{x} = \int_0^h \left(\int_{\mathcal{Q}_t} dx_1 \cdots dx_{n-1}\right) dx_n \\ &= \int_0^h \text{vol}\,(\mathcal{Q}_t)dt = \int_0^h \text{vol}\,(\mathcal{Q}_0)dt = \text{vol}\,(\mathcal{Q})h.\end{aligned}$$

Now we can prove Theorem B.1.

Proof of Theorem B.1

We prove by induction on n. The $n = 1$ case is obvious. Assume that we have proved the $n-1$ case. Now given that $\mathscr{P}$ is a parallelepiped in $\mathbb{R}^n$ which is spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_n$, we can assume that $\mathscr{P}$ is based at the origin $\mathbf{0}$ because translations preserve volumes. Let

$$A = \left[\mathbf{v}_1 \,\middle|\, \cdots \,\middle|\, \mathbf{v}_n\right].$$

We want to show that

$$\text{vol}\,(\mathscr{P}) = |\det A|.$$

Let W be the subspace of $\mathbb{R}^{n-1}$ that is spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_{n-1}$. Applying the Gram-Schmidt process to the basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ of $\mathbb{R}^n$, we obtain an orthonormal basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$. By the algorithm, the unit vector $\mathbf{u}_n$ is orthogonal to the subspace W. Let

$$P = \left[\, \mathbf{u}_1 \;\middle|\; \cdots \;\middle|\; \mathbf{u}_n \,\right]$$

be the orthogonal matrix whose column vectors are $\mathbf{u}_1, \ldots, \mathbf{u}_n$, and consider the orthogonal transformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{T}(\mathbf{x}) = P^{-1}\mathbf{x} = P^T\mathbf{x}$. For $1 \le i \le n$, let

$$\widetilde{\mathbf{v}}_i = \mathbf{T}(\mathbf{v}_i).$$

Then $\widetilde{\mathscr{P}} = \mathbf{T}(\mathscr{P})$ is a parallelepiped that has the same volume as $\mathscr{P}$, and it is spanned by $\widetilde{\mathbf{v}}_1, \ldots, \widetilde{\mathbf{v}}_n$. Notice that

$$\begin{aligned}
\widetilde{A} &= \left[\, \widetilde{\mathbf{v}}_1 \;\middle|\; \cdots \;\middle|\; \widetilde{\mathbf{v}}_n \,\right] = P^T \left[\, \mathbf{v}_1 \;\middle|\; \cdots \;\middle|\; \mathbf{v}_n \,\right] \\
&= \left[\begin{array}{c} \mathbf{u}_1^T \\ \hline \vdots \\ \hline \mathbf{u}_n^T \end{array}\right] \left[\, \mathbf{v}_1 \;\middle|\; \cdots \;\middle|\; \mathbf{v}_n \,\right] = \left[\begin{array}{c|c} & \langle \mathbf{u}_1, \mathbf{v}_n \rangle \\ B & \vdots \\ & \langle \mathbf{u}_{n-1}, \mathbf{v}_n \rangle \\ \hline \mathbf{0} & \langle \mathbf{u}_n, \mathbf{v}_n \rangle \end{array}\right].
\end{aligned}$$

From this, we find that

$$\det(\widetilde{A}) = \det(B) \times \langle \mathbf{u}_n, \mathbf{v}_n \rangle.$$

Comparing the columns, we also have

$$\widetilde{\mathbf{v}}_i = (\mathbf{z}_i, 0) \qquad \text{for } 1 \le i \le n-1,$$

where $\mathbf{z}_1, \ldots, \mathbf{z}_{n-1}$ are the column vectors of B, which are vectors in $\mathbb{R}^{n-1}$; and

$$\widetilde{\mathbf{v}}_n = \left[\begin{array}{c} \langle \mathbf{u}_1, \mathbf{v}_n \rangle \\ \vdots \\ \langle \mathbf{u}_{n-1}, \mathbf{v}_n \rangle \\ \hline \langle \mathbf{u}_n, \mathbf{v}_n \rangle \end{array}\right].$$

The transformation $\mathbf{T}$ maps the subspace W to the hyperplane $x_n = 0$, which can be identified with $\mathbb{R}^{n-1}$. Let $\mathcal{Q}$ be the parallelepiped in $\mathbb{R}^{n-1}$ based at the origin and spanned by the vectors $\mathbf{z}_1, \ldots, \mathbf{z}_{n-1}$. The volume of the parallepiped $\widetilde{\mathscr{P}}$ is equal to the volume of $\mathcal{Q}$ times the distance h from the tip of the vector $\widetilde{\mathbf{v}}_n$ to the plane $x_n = 0$. By definition,

$$h = \left\| \operatorname{proj}_{\mathbf{e}_n} \widetilde{\mathbf{v}}_n \right\| = \left| \langle \mathbf{u}_n, \mathbf{v}_n \rangle \right| .$$

Proposition B.3 gives

$$\operatorname{vol}\left(\widetilde{\mathscr{P}}\right) = \operatorname{vol}\left(\mathcal{Q}\right) \times \left| \langle \mathbf{u}_n, \mathbf{v}_n \rangle \right| .$$

By inductive hypothesis,

$$\operatorname{vol}\left(\mathcal{Q}\right) = |\det(B)|.$$

Therefore,

$$\operatorname{vol}\left(\widetilde{\mathscr{P}}\right) = \left|\det(B) \times \langle \mathbf{u}_n, \mathbf{v}_n \rangle\right| = |\det(\widetilde{A})|.$$

Since $\widetilde{A} = P^T A$, we find that

$$\det(\widetilde{A}) = \det(P^T)\det(A) = \pm \det(A).$$

Hence,

$$\operatorname{vol}\left(\mathscr{P}\right) = \operatorname{vol}\left(\widetilde{\mathscr{P}}\right) = |\det(A)|.$$

This completes the proof of Theorem B.1.

As a corollary , we have the following.

Theorem B.4

Let $\mathbf{I}$ be a closed rectangle in $\mathbb{R}^n$, and let $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{T}(\mathbf{x}) = A\mathbf{x}$ be an invertible linear transformation. Then

$$\operatorname{vol}\left(\mathbf{T}(\mathbf{I})\right) = |\det A| \operatorname{vol}\left(\mathbf{I}\right).$$

Proof

Let $\mathbf{I} = \prod_{i=1}^{n}[a_i, b_i]$. Then $\mathbf{I} = \mathbf{S}(Q_n) + \mathbf{a}$, where $\mathbf{a} = (a_1, \ldots, a_n)$, Q_n is the standard unit cube $[0, 1]^n$, and $\mathbf{S} : \mathbb{R}^n \to \mathbb{R}^n$ is the linear transformation defined by the diagonal matrix B with diagonal entries $b_1 - a_1, b_2 - a_2, \ldots, b_n - a_n$. Therefore,

$$\mathbf{T}(\mathbf{I}) = (\mathbf{T} \circ \mathbf{S})(Q_n) + \mathbf{T}(\mathbf{a}).$$

Since the matrix associated with the linear transformation $(\mathbf{T} \circ \mathbf{S}) : \mathbb{R}^n \to \mathbb{R}^n$ is AB, $\mathbf{T}(\mathbf{I})$ is a parallelepiped based at $\mathbf{T}(\mathbf{a})$ and spanned by the columnn vectors of AB. By Theorem B.1,

$$\text{vol } (\mathbf{T}(\mathbf{I})) = |\det(AB)| = |\det(A)||\det(B)|.$$

Obviously,

$$|\det B| = \prod_{i=1}^{n}(b_i - a_i) = \text{vol } (\mathbf{I}).$$

This proves that

$$\text{vol } (\mathbf{T}(\mathbf{I})) = |\det A| \text{ vol } (\mathbf{I}).$$

Remark B.2

The formula

$$\text{vol } (\mathbf{T}(\mathbf{I})) = |\det A| \text{ vol } (\mathbf{I})$$

still holds even though the matrix A is not invertible. In this case, $\det A = 0$, and the column vectors of A are not linearly independent. Therefore, $\mathbf{T}(\mathbf{I})$ lies in a plane in $\mathbb{R}^n$, and so $\mathbf{T}(\mathbf{I})$ has zero volume.

Appendix C

Necessary and Sufficient Condition for Riemann Integrability

In this appendix, we want to prove the Lebesgue-Vitali theorem which gives a necessary and sufficient condition for a bounded function $f : \mathfrak{D} \to \mathbb{R}$ to be Riemann integrable. We will introduce the concept of Lebesgue measure zero without introducing the concept of general Lebesgue measure. The latter is often covered in a standard course in real analysis.

Recall that the volume of a closed rectangle $\mathbf{I} = \prod_{i=1}^{n}[a_i, b_i]$ or its interior $\text{int}\,(\mathbf{I}) = \prod_{i=1}^{n}(a_i, b_i)$ is

$$\text{vol}\,(\mathbf{I}) = \text{vol}\,(\text{int}\,\mathbf{I}) = \prod_{i=1}^{n}(b_i - a_i).$$

If A is a subset of $\mathbb{R}^n$, we say that A has Jordan content zero if

(i) for every $\varepsilon > 0$, there are finitely many closed rectangles $\mathbf{I}_1, \ldots, \mathbf{I}_k$ such that

$$A \subset \bigcup_{j=1}^{k} \mathbf{I}_j \quad \text{and} \quad \sum_{j=1}^{k} \text{vol}\,(\mathbf{I}_j) < \varepsilon.$$

This is equivalent to any of the followings.

(ii) For every $\varepsilon > 0$, there are finitely many closed cubes $Q_1, \ldots, Q_k$ such that

$$A \subset \bigcup_{j=1}^{k} Q_j \quad \text{and} \quad \sum_{j=1}^{k} \text{vol}\,(Q_j) < \varepsilon.$$

(iii) For every $\varepsilon > 0$, there are finitely many open rectangles $U_1, \ldots, U_k$ such that

$$A \subset \bigcup_{j=1}^{k} U_j \quad \text{and} \quad \sum_{j=1}^{k} \text{vol}\,(U_j) < \varepsilon.$$

(iv) For every $\varepsilon > 0$, there are finitely many open cubes $V_1, \ldots, V_k$ such that

$$A \subset \bigcup_{j=1}^{k} V_j \quad \text{and} \quad \sum_{j=1}^{k} \text{vol}\,(V_j) < \varepsilon.$$

A set has Jordan content zero if and only if it is Jordan measurable and its volume is zero. Hence, we also call a set that has Jordan content zero as a set that has *Jordan measure zero*. The Jordan measure of a Jordan measurable set A is the volume of A defined as the Riemann integral of the characteristic function $\chi_A : A \to \mathbb{R}$.

In Lebesgue measure, instead of a covering by finitely many rectangles, we allow a covering by countably many rectangles. A set S is countable if it is finite or it is countably infinite. The latter means that there is a one-to-one correspondence between S and the set $\mathbb{Z}^+$. In any case, a set S is countable if and only if there is a surjection $h : \mathbb{Z}^+ \to S$, which allows us to write

$$S = \left\{ s_k \,|\, k \in \mathbb{Z}^+ \right\}, \qquad \text{where } s_k = h(k).$$

Definition C.1 Lebesgue Measure Zero

Let A be a subset of $\mathbb{R}^n$. We say that A has Lebesgue measure zero if for every $\varepsilon > 0$, there is a countable collection of open rectangles $\{U_k \,|\, k \in \mathbb{Z}^+\}$ that covers A, the sum of whose volumes is less than ε. Namely,

$$A \subset \bigcup_{k=1}^{\infty} U_k \quad \text{and} \quad \sum_{k=1}^{\infty} \text{vol}\,(U_k) < \varepsilon.$$

The following is obvious.

Proposition C.1

Let A be a subset of $\mathbb{R}^n$. If A has Jordan content zero, then it has Lebesgue measure zero.

The converse is not true. There are sets with Lebesgue measure zero, but they do not have Jordan content zero. The following gives an example of such sets.

Example C.1

Let $A = \mathbb{Q} \cap [0,1]$. The function $\chi_A : [0,1] \to \mathbb{R}$ is the Dirichlet's function, which is not Riemann integrable. Hence, A is not Jordan measurable. Nevertheless, we claim that A has Lebesgue measure zero.
Recall that $\mathbb{Q}$ is a countable set. As a subset of $\mathbb{Q}$, A is also countable. Hence, we can write A as

$$A = \{a_k \,|\, k \in \mathbb{Z}^+\}.$$

Given $\varepsilon > 0$ and $k \in \mathbb{Z}^+$, let U_k be the open rectangle

$$U_k = \left(a_k - \frac{\varepsilon}{2^{k+2}}, a_k + \frac{\varepsilon}{2^{k+2}}\right).$$

Then $a_k \in U_k$ for each $k \in \mathbb{Z}^+$. Thus,

$$A \subset \bigcup_{k=1}^{\infty} U_k.$$

Now,

$$\sum_{k=1}^{\infty} \text{vol}\,(U_k) = \sum_{k=1}^{\infty} \frac{\varepsilon}{2^{k+1}} = \frac{\varepsilon}{2} < \varepsilon.$$

Therefore, A has Lebesgue measure zero.

The converse to Proposition C.1 is true if A is compact.

Proposition C.2

Let A be a compact subset of $\mathbb{R}^n$. If A has Lebesgue measure zero, then it has Jordan content zero.

Proof

Given $\varepsilon > 0$, since A has Lebesgue measure zero, there is a countable collection $\{U_\alpha \,|\, \alpha \in \mathbb{Z}^+\}$ of open rectangles that covers A, and

$$\sum_{\alpha \in \mathbb{Z}^+} \text{vol}\,(U_\alpha) < \varepsilon.$$

Since A is compact, there is a finite subcollection $\{U_{\alpha_l} \mid 1 \leq l \leq m\}$ that covers A. Obviously, we also have

$$\sum_{l=1}^{m} \text{vol}\,(U_{\alpha_l}) < \varepsilon.$$

Hence, A has Jordan content 0.

Example C.2

Using the same reasoning as in Example C.1, one can show that any countable subset of $\mathbb{R}^n$ has Lebesgue measure zero.

We have seen that if A is a subset of $\mathbb{R}^n$ that has Jordan content zero, then its closure $\overline{A}$ also has Jordan content zero. However, the same is not true for Lebesgue measure.

Example C.3

Example C.1 shows that the set $A = \mathbb{Q} \cap [0, 1]$ has Lebesgue measure zero. Notice that $\overline{A} = [0, 1]$. It cannot have Lebesgue measure zero.

As in the case of Jordan content zero, we have the following equivalences for a set A in $\mathbb{R}^n$ to have Lebesgue measure zero.

(i) For every $\varepsilon > 0$, there is a countable collection of open rectangles $\{U_k \mid k \in \mathbb{Z}^+\}$ such that

$$A \subset \bigcup_{k=1}^{\infty} U_k \quad \text{and} \quad \sum_{k=1}^{\infty} \text{vol}\,(U_k) < \varepsilon.$$

(ii) For every $\varepsilon > 0$, there is a countable collection of closed rectangles $\{\mathbf{I}_k \mid k \in \mathbb{Z}^+\}$ such that

$$A \subset \bigcup_{k=1}^{\infty} \mathbf{I}_k \quad \text{and} \quad \sum_{k=1}^{\infty} \text{vol}\,(\mathbf{I}_k) < \varepsilon.$$

(iii) For every $\varepsilon > 0$, there is a countable collection of open cubes $\{V_k \mid k \in \mathbb{Z}^+\}$

such that

$$A \subset \bigcup_{k=1}^{\infty} V_k \quad \text{and} \quad \sum_{k=1}^{\infty} \text{vol}\,(V_k) < \varepsilon.$$

(iv) For every $\varepsilon > 0$, there is a countable collection of closed cubes $\{Q_k \,|\, k \in \mathbb{Z}^+\}$ such that

$$A \subset \bigcup_{k=1}^{\infty} Q_k \quad \text{and} \quad \sum_{k=1}^{\infty} \text{vol}\,(Q_k) < \varepsilon.$$

The following is obvious.

Proposition C.3

Let A be a subset of $\mathbb{R}^n$. If A has Lebesgue measure zero, and B is a subset of A, then B also has Lebesgue measure zero.

Using the fact that the set $\mathbb{Z}^+ \times \mathbb{Z}^+$ is countable, we find that a countable union of coutable sets is countable. This gives the following.

Proposition C.4

Let $\{A_m \,|\, m \in \mathbb{Z}^+\}$ be a countable collection of subsets of $\mathbb{R}^n$. If each of the $A_m, m \in \mathbb{Z}^+$ has Lebesgue measure zero, then the set $A = \bigcup_{m=1}^{\infty} A_m$ also has Lebesgue measure zero.

Proof

Fixed $\varepsilon > 0$. For each $m \in \mathbb{Z}^+$, since A_m has Lebesgue measure zero, there is a countable collection $\mathscr{B}_m = \{U_{m,k} \,|\, k \in \mathbb{Z}^+\}$ of open rectangles such that

$$A_m \subset \bigcup_{k=1}^{\infty} U_{m,k} \quad \text{and} \quad \sum_{k=1}^{\infty} \text{vol}\,(U_{m,k}) < \frac{\varepsilon}{2^m}.$$

It follows that

$$A \subset \bigcup_{m=1}^{\infty} \bigcup_{k=1}^{\infty} U_{m,k} \quad \text{and} \quad \sum_{m=1}^{\infty} \sum_{k=1}^{\infty} \text{vol}\,(U_{m,k}) < \sum_{m=1}^{\infty} \frac{\varepsilon}{2^m} = \varepsilon.$$

Notice that the collection

$$\mathscr{B} = \bigcup_{m=1}^{\infty} \mathscr{B}_m = \left\{ U_{m,k} \,|\, m \times k \in \mathbb{Z}^+ \times \mathbb{Z}^+ \right\}$$

is countable. This proves that A has Lebesgue measure zero.

Now we proceed to the main theorem.

Theorem C.5 Lebesgue-Vitali Theorem

Let $\mathbf{I} = \prod_{i=1}^{n} [a_i, b_i]$ be a closed rectangle in $\mathbb{R}^n$. Given a bounded function $f : \mathbf{I} \to \mathbb{R}$, let $\mathcal{N}$ be its set of discontinuities. Then $f : \mathbf{I} \to \mathbb{R}$ is Riemann integrable if and only if $\mathcal{N}$ has Lebesgue measure zero.

In this theorem, we only consider functions defined on closed rectangles. This is because the Riemann integrability of a function $f : \mathfrak{D} \to \mathbb{R}$ defined on a bounded set $\mathfrak{D}$ is defined in terms of the Riemann integrability of its zero extension $\check{f} : \mathbf{I} \to \mathbb{R}$ to a closed rectangle $\mathbf{I}$ that contains $\mathfrak{D}$.

To prove the Lebesgue-Vitali theorem, we need a few lemmas. Given a bounded function $f : \mathfrak{D} \to \mathbb{R}$, we define the oscillation of f at a point $\mathbf{x}_0 \in \mathfrak{D}$ as

$$\omega_f(\mathbf{x}_0) = \lim_{r \to 0^+} \sup \left\{ f(\mathbf{u}) - f(\mathbf{v}) \,|\, \mathbf{u}, \mathbf{v} \in B(\mathbf{x}_0, r) \cap \mathfrak{D} \right\}.$$

Notice that the set

$$\mathcal{F}_r = \left\{ f(\mathbf{u}) - f(\mathbf{v}) \,|\, \mathbf{u}, \mathbf{v} \in B(\mathbf{x}_0, r) \cap \mathfrak{D} \right\}$$

is a bounded subset of real numbers and $-\mathcal{F}_r = \mathcal{F}_r$. Thus, the supremum of $\mathcal{F}_r$ always exists and is nonnegative. It is easy to see that

$$\mathcal{F}_{r_1} \subset \mathcal{F}_{r_2} \qquad \text{if } r_1 < r_2.$$

Therefore, $\sup \mathcal{F}_r$ decreases as $r \to 0^+$ This implies that

$$\omega_f(\mathbf{x}_0) = \lim_{r \to 0^+} \sup \mathcal{F}_r = \inf_{r > 0} \sup_{\mathbf{u}, \mathbf{v} \in B(\mathbf{x}_0, r) \cap \mathfrak{D}} (f(\mathbf{u}) - f(\mathbf{v}))$$

exists and is nonnegative.

Lemma C.6

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$, and let $\mathbf{x}_0$ be a point in $\mathfrak{D}$. Assume that $f : \mathfrak{D} \to \mathbb{R}$ is a bounded function. Then f is continuous at $\mathbf{x}_0$ if and only if $\omega_f(\mathbf{x}_0) = 0$.

Proof

First assume that f is continuous at $\mathbf{x}_0$. Given $\varepsilon > 0$, there is a $\delta > 0$ such that for all $\mathbf{x} \in B(\mathbf{x}_0, \delta) \cap \mathfrak{D}$,

$$|f(\mathbf{x}) - f(\mathbf{x}_0)| < \frac{\varepsilon}{3}.$$

It follows that for all $\mathbf{u}, \mathbf{v} \in B(\mathbf{x}_0, \delta) \cap \mathfrak{D}$,

$$|f(\mathbf{u}) - f(\mathbf{v})| < \frac{2\varepsilon}{3}.$$

Thus, if $r < \delta$,

$$0 \leq \sup\{f(\mathbf{u}) - f(\mathbf{v}) \,|\, \mathbf{u}, \mathbf{v} \in B(\mathbf{x}_0, r) \cap \mathfrak{D}\} \leq \frac{2\varepsilon}{3} < \varepsilon.$$

This shows that

$$\omega_f(\mathbf{x}_0) = \lim_{r \to 0^+} \sup\{f(\mathbf{u}) - f(\mathbf{v}) \,|\, \mathbf{u}, \mathbf{v} \in B(\mathbf{x}_0, r) \cap \mathfrak{D}\} = 0.$$

Conversely, assume that $\omega_f(\mathbf{x}_0) = 0$. Given $\varepsilon > 0$, there is a $\delta > 0$ such that for all $0 < r < \delta$,

$$\sup\{f(\mathbf{u}) - f(\mathbf{v}) \,|\, \mathbf{u}, \mathbf{v} \in B(\mathbf{x}_0, r) \cap \mathfrak{D}\} < \varepsilon.$$

If $\mathbf{x}$ is in $B(\mathbf{x}_0, \delta/2) \cap \mathfrak{D}$,

$$|f(\mathbf{x}) - f(\mathbf{x}_0)| \leq \sup\{f(\mathbf{u}) - f(\mathbf{v}) \,|\, \mathbf{u}, \mathbf{v} \in B(\mathbf{x}_0, \delta/2) \cap \mathfrak{D}\} < \varepsilon.$$

This proves that f is continuous at $\mathbf{x}_0$.

Corollary C.7

Let $\mathfrak{D}$ be a subset of $\mathbb{R}^n$, and let $f : \mathfrak{D} \to \mathbb{R}$ be a bounded function defined on $\mathfrak{D}$. If $\mathcal{N}$ is the set of discontinuities of f, then

$$\mathcal{N} = \{\mathbf{x} \in \mathfrak{D} \mid \omega_f(\mathbf{x}) > 0\}.$$

We also need the following proposition.

Proposition C.8

Let $\mathfrak{D}$ be a compact subset of $\mathbb{R}^n$, and let $f : \mathfrak{D} \to \mathbb{R}$ be a bounded function defined on $\mathfrak{D}$.

(a) For any $a > 0$, the set

$$A = \{\mathbf{x} \in \mathfrak{D} \mid \omega_f(\mathbf{x}) \geq a\}$$

is a compact subset of $\mathbb{R}^n$.

(b) If $\mathcal{N}$ is the set of discontinuities of f, then

$$\mathcal{N} = \bigcup_{k=1}^{\infty} \mathcal{N}_k,$$

where

$$\mathcal{N}_k = \left\{ \mathbf{x} \in \mathfrak{D} \;\middle|\; \omega_f(\mathbf{x}) \geq \frac{1}{k} \right\}.$$

(c) The set $\mathcal{N}$ has Lebesgue measure zero if and only if $\mathcal{N}_k$ has Jordan content zero for each $k \in \mathbb{Z}^+$.

Proof

Since $\mathfrak{D}$ is compact, it is closed and bounded. For part (a), $A \subset \mathfrak{D}$ implies A is bounded. To prove that A is compact, we only need to show that A is closed. This is equivalent to $\mathbb{R}^n \setminus A$ is open. Notice that

$$\mathbb{R}^n \setminus A = U_1 \cup U_2,$$

where

$$U_1 = \mathbb{R}^n \setminus \mathfrak{D} \quad \text{and} \quad U_2 = \mathfrak{D} \setminus A.$$

Since $\mathfrak{D}$ is closed, U_1 is open. If $\mathbf{x}_0 \in U_2$, then $\omega_f(\mathbf{x}_0) < a$. Let $\varepsilon = a - \omega_f(\mathbf{x}_0)$. Then $\varepsilon > 0$. By definition of $\omega_f(\mathbf{x}_0)$, there is a $\delta > 0$ such that for all $0 < r < \delta$,

$$\sup\{f(\mathbf{u}) - f(\mathbf{v}) \,|\, \mathbf{u}, \mathbf{v} \in B(\mathbf{x}_0, r) \cap \mathfrak{D}\} < \omega_f(\mathbf{x}_0) + \varepsilon = a.$$

Take $c = \delta/3$. If $\mathbf{x}$ is in $B(\mathbf{x}_0, c)$, and $\mathbf{u}, \mathbf{v}$ are in $B(\mathbf{x}, c)$, then $\mathbf{u}, \mathbf{v}$ are in $B(\mathbf{x}_0, 2c)$. Since $2c < \delta$, we find that

$$\begin{aligned} \omega_f(\mathbf{x}) &\leq \sup\{f(\mathbf{u}) - f(\mathbf{v}) \,|\, \mathbf{u}, \mathbf{v} \in B(\mathbf{x}, c) \cap \mathfrak{D}\} \\ &\leq \sup\{f(\mathbf{u}) - f(\mathbf{v}) \,|\, \mathbf{u}, \mathbf{v} \in B(\mathbf{x}_0, 2c) \cap \mathfrak{D}\} < a. \end{aligned}$$

This shows that $B(\mathbf{x}_0, c) \subset U_2$. Hence, U_2 is open. Since $\mathbb{R}^n \setminus A$ is a union of two open sets, it is open. This completes the proof.
Part (b) follows from Corollary C.7 and the identity

$$(0, \infty) = \bigcup_{k=1}^{\infty} \left[\frac{1}{k}, \infty\right).$$

For part (c), if $\mathcal{N}$ has Lebesgue measure zero, then for any $k \in \mathbb{Z}^+$, $\mathcal{N}_k$ also has Lebesgue measure zero. By part (a) and Proposition C.2, $\mathcal{N}_k$ has Jordan content zero. Conversely, assume that $\mathcal{N}_k$ has Jordan content zero for each $k \in \mathbb{Z}^+$. Then $\mathcal{N}_k$ has Lebesgue measure zero for each $k \in \mathbb{Z}^+$. Part (b) and Proposition C.4 implies that $\mathcal{N}$ also has Lebesgue measure zero.

Now we can prove the Lebesgue-Vitali theorem.

Proof of the Lebesgue-Vitali Theorem

First we assume that $f : \mathbf{I} \to \mathbb{R}$ is Riemann integrable. Given $k \in \mathbb{Z}^+$, we will show that the set

$$\mathcal{N}_k = \left\{ \mathbf{x} \in \mathfrak{D} \;\middle|\; \omega_f(\mathbf{x}) \geq \frac{1}{k} \right\}$$

has Jordan content zero. By Proposition C.8, this implies that the set $\mathcal{N}$ of discontiuities of $f : \mathbf{I} \to \mathbb{R}$ has Lebesgue measure zero.
Fixed $k \in \mathbb{Z}^+$. Given $\varepsilon > 0$, since $f : \mathbf{I} \to \mathbb{R}$ is Riemann integrable, there is a partition $\mathbf{P}$ of $\mathbf{I}$ such that

$$U(f, \mathbf{P}) - L(f, \mathbf{P}) < \frac{\varepsilon}{2k}.$$

Let

$$\mathscr{A} = \{\mathbf{J} \in \mathcal{J}_\mathbf{P} \,|\, (\operatorname{Int} \mathbf{J}) \cap \mathcal{N}_k \neq \emptyset\}.$$

Then

$$\mathcal{N}_k = A_1 \cup A_2,$$

where

$$A_1 = \left(\bigcup_{\mathbf{J} \in \mathscr{A}} \operatorname{int} \mathbf{J} \right) \cap \mathcal{N}_k,$$

and

$$A_2 = \left(\bigcup_{\mathbf{J} \in \mathcal{J}_\mathbf{P}} \partial \mathbf{J} \right) \cap \mathcal{N}_k.$$

Notice that the set A_2 has Jordan content zero. Therefore, there are finitely many open rectangles $U_1, \ldots, U_m$ such that

$$A_2 \subset \bigcup_{l=1}^{m} U_l \quad \text{and} \quad \sum_{l=1}^{m} \operatorname{vol}(U) < \frac{\varepsilon}{2}.$$

The set A_1 itself is contained in a finite union of open rectangles $\operatorname{int} \mathbf{J}$ with $\mathbf{J} \in \mathscr{A}$. Notice that

$$\sum_{\mathbf{J} \in \mathscr{A}} (M_\mathbf{J}(f) - m_\mathbf{J}(f)) \operatorname{vol}(\mathbf{J}) \leq U(f, \mathbf{P}) - L(f, \mathbf{P}) < \frac{\varepsilon}{2k}.$$

If $\mathbf{J}$ is in $\mathscr{A}$, there is an $\mathbf{x}_0 \in A_1$ such that $\mathbf{x}_0 \in \operatorname{int} \mathbf{J}$. Since $\operatorname{int} \mathbf{J}$ is an open set, there is a $\delta > 0$ such that $B(\mathbf{x}_0, \delta) \subset \mathbf{J}$.

Now,

$$\sup\{f(\mathbf{u}) - f(\mathbf{v}) \,|\, \mathbf{u}, \mathbf{v} \in B(\mathbf{x}_0, \delta)\} \geq \omega_f(\mathbf{x}_0) \geq \frac{1}{k}.$$

Therefore,

$$M_{\mathbf{J}}(f) - m_{\mathbf{J}}(f) \geq \frac{1}{k}$$

for each $\mathbf{J}$ in $\mathscr{A}$. This implies that

$$\frac{1}{k} \sum_{\mathbf{J} \in \mathscr{A}} \text{vol}\,(\text{int}\,\mathbf{J}) \leq \sum_{\mathbf{J} \in \mathscr{A}} (M_{\mathbf{J}}(f) - m_{\mathbf{J}}(f))\,\text{vol}\,(\text{int}\,\mathbf{J}) < \frac{\varepsilon}{2k}.$$

Thus,

$$\sum_{\mathbf{J} \in \mathscr{A}} \text{vol}\,(\text{int}\,\mathbf{J}) < \frac{\varepsilon}{2}.$$

Hence,

$$\mathscr{B} = \{\text{int}\,\mathbf{J} \,|\, \mathbf{J} \in \mathscr{A}\} \cup \{U_l \,|\, 1 \leq l \leq m\}$$

is a finite collection of open rectangles that covers $\mathcal{N}_k$, and the sum of the volumes of the rectangles in $\mathscr{B}$ is less than ε. This proves that $\mathcal{N}_k$ indeed has Jordan content zero.

Conversely, assume that $\mathcal{N}$ has Lebesgue measure zero. Since $f : \mathbf{I} \to \mathbb{R}$ is bounded, there is a positive number M such that

$$|f(\mathbf{x})| \leq M \qquad \text{for all } \mathbf{x} \in \mathbf{I}.$$

Given $\varepsilon > 0$, let k be a positive integer such that

$$k \geq \frac{2\text{vol}\,(\mathbf{I})}{\varepsilon}.$$

Proposition C.8 says that $\mathcal{N}_k$ has Jordan content zero. Thus, there is a finite collection of open rectangles $\mathscr{B}_k = \{U_l \,|\, 1 \leq l \leq m\}$ such that

$$\mathcal{N}_k \subset \bigcup_{l=1}^{m} U_l \qquad\qquad \sum_{l=1}^{m} \text{vol}\,(U_l) < \frac{\varepsilon}{4M}.$$

Let $\mathbf{P}_0$ be a partition of $\mathbf{I}$ such that each rectangle $\mathbf{J}$ in $\mathcal{J}_{\mathbf{P}_0}$ lies entirely in the closure of one of the rectangles U_l, $1 \leq l \leq m$ or it is disjoint from all the U_l, $1 \leq l \leq m$. Let

$$\mathscr{C} = \{\mathbf{J} \in \mathcal{J}_{\mathbf{P}_0} \,|\, \mathbf{J} \cap U_l = \emptyset \text{ for all } 1 \leq l \leq m\}.$$

Then

$$\mathcal{N}_k \subset \bigcup_{\mathbf{J}\in\mathcal{J}_{\mathbf{P}_0}\setminus\mathscr{C}} \operatorname{int}\mathbf{J}.$$

For each point $\mathbf{x}$ that is in $\mathbf{I}\setminus\mathcal{N}_k$, there is an $r_{\mathbf{x}} > 0$ such that the open cube $\mathbf{x} + (-r_{\mathbf{x}}, r_{\mathbf{x}})^n$ is contained in the open set $\mathbb{R}\setminus\mathcal{N}_k$. By taking a smaller $r_{\mathbf{x}}$, we can assume that

$$\sup\{f(\mathbf{u}) - f(\mathbf{v}) \,|\, \mathbf{u},\mathbf{v}\in B(\mathbf{x}, 2r_{\mathbf{x}})\} < \frac{1}{k}.$$

The ball $B(\mathbf{x}, 2r_{\mathbf{x}})$ contains the cube $Q = \mathbf{x} + [-r_{\mathbf{x}}, r_{\mathbf{x}}]^n$. Therefore,

$$M_Q(f) - m_Q(f) \le \frac{1}{k}.$$

The collection

$$\{\mathbf{x} + (-r_{\mathbf{x}}, r_{\mathbf{x}})^n \,|\, \mathbf{x}\in\mathbf{I}\setminus\mathcal{N}_k\}$$

is an open covering of the compact set

$$K = \bigcup_{\mathbf{J}\in\mathscr{C}} \mathbf{J}.$$

Thus, there is a finite subcover $\{V_1,\ldots,V_s\}$. For $1\le j\le s$, let $\mathbf{I}_j = \overline{V}_j\cap\mathbf{I}$. Then we still have

$$K = \bigcup_{\mathbf{J}\in\mathscr{C}} \mathbf{J} \subset \bigcup_{j=1}^{s} \mathbf{I}_j.$$

After renaming the rectangles, let

$$\left\{\overline{U}_l \,|\, 1\le l\le m\right\} \cup \left\{\mathbf{I}_j \,|\, 1\le j\le s\right\} = \left\{W_1, W_2, \ldots, W_q\right\}.$$

Now let $\mathbf{P}$ be a partition of $\mathbf{I}$ so that each rectangle in $\mathcal{J}_{\mathbf{P}}$ is either disjoint from the interior of all the W_j, $1\le j\le q$, or is contained in one of the W_j. Let

$$\mathscr{D} = \left\{\mathbf{J}\in\mathcal{J}_{\mathcal{P}} \,|\, \mathbf{J}\in\overline{U}_l \text{ for some } 1\le l\le m\right\}.$$

Then

$$\sum_{\mathbf{J}\in\mathscr{D}} (M_{\mathbf{J}} - m_{\mathbf{J}})\,\mathrm{vol}\,(\mathbf{J}) \le 2M\sum_{l=1}^{m} \mathrm{vol}\,\left(\overline{U}_l\right) = 2M\sum_{l=1}^{m} \mathrm{vol}\,\left(U_l\right) < \frac{\varepsilon}{2}.$$

For those $\mathbf{J}$ that is not in $\mathscr{D}$, it is contained in one of the cubes $\mathbf{x}+[-r_{\mathbf{x}}, r_{\mathbf{x}}]^n$. Therefore,

$$M_{\mathbf{J}}(f) - m_{\mathbf{J}}(f) \leq \frac{1}{k}.$$

It follows that

$$\sum_{\mathbf{J}\in\mathcal{J}_{\mathbf{P}}\setminus\mathscr{D}} (M_{\mathbf{J}} - m_{\mathbf{J}}) \operatorname{vol}(\mathbf{J}) \leq \frac{1}{k} \sum_{\mathbf{J}\in\mathcal{J}_{\mathbf{P}}\setminus\mathscr{D}} \operatorname{vol}(\mathbf{J}) \leq \frac{\operatorname{vol}(\mathbf{I})}{k} \leq \frac{\varepsilon}{2}.$$

This proves that

$$U(f, \mathbf{P}) - L(f, \mathbf{P}) < \varepsilon.$$

Hence, $f : \mathbf{I} \to \mathbb{R}$ is Riemann integrable.

References

[Abb15] Stephen Abbott, *Understanding analysis*, second ed., Undergraduate Texts in Mathematics, Springer, New York, 2015. MR 3331079

[Apo74] Tom M. Apostol, *Mathematical analysis*, second ed., Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont., 1974. MR 0344384

[BS92] Robert G. Bartle and Donald R. Sherbert, *Introduction to real analysis*, second ed., John Wiley & Sons, Inc., New York, 1992. MR 1135107

[CB84] Ruel V. Churchill and James Ward Brown, *Complex variables and applications*, fourth ed., McGraw-Hill Book Co., New York, 1984. MR 730937

[Fit09] Patrick M. Fitzpatrick, *Advanced calculus*, second ed., American Mathematical Society, 2009.

[Rud76] Walter Rudin, *Principles of mathematical analysis*, third ed., International Series in Pure and Applied Mathematics, McGraw-Hill Book Co., New York-Auckland-Düsseldorf, 1976. MR 0385023

[SCW20] James Stewart, Daniel K. Clegg, and Saleem Watson, *Calculus*, ninth ed., Cengage Learning, 2020.

[SS03] Elias M. Stein and Rami Shakarchi, *Fourier analysis*, Princeton Lectures in Analysis, vol. 1, Princeton University Press, Princeton, NJ, 2003, An introduction. MR 1970295

[Tao14] Terence Tao, *Analysis. II*, third ed., Texts and Readings in Mathematics, vol. 38, Hindustan Book Agency, New Delhi, 2014. MR 3310023

[Tao16] ———, *Analysis. I*, third ed., Texts and Readings in Mathematics, vol. 37, Hindustan Book Agency, New Delhi; Springer, Singapore, 2016, Edectronic edition of [MR3309891]. MR 3728289

[Zor15] Vladimir A. Zorich, *Mathematical analysis. I*, second ed., Universitext, Springer-Verlag, Berlin, 2015, With Appendices A–F and new problems translated by Octavio Paniagua T. MR 3495809

[Zor16] ———, *Mathematical analysis. II*, second ed., Universitext, Springer, Heidelberg, 2016. MR 3445604