Math412 Notes

Analysis of Functions of Several Variables

Math412 Notes

Analysis of Functions of Several Variables

Zheng-Chao Han

February 13, 2024

Preface

These notes have been constructed to cover the topics in Math412. While there are many excellent textbooks covering the analysis of differential and integral calculus in a single variable---the bulk of the topics for Math411, there are few choices covering the integration of functions and differential forms in several variables that provide more geometric discussions and motivation for some of the more abstract notions related to differential forms. This was the initial motivation for constructing these notes. But the notes have been expanded to cover aspects of analysis of a single variable which often form a bridging part with the first semester course mostly surrounding the notions of pointwise convergence and uniform convergence.

Among the most basic questions of this part are

- If a sequence of continuous functions f_k on $[a, b]$ converges pointwise to a function f, what can be said about f? Is f necessarily continuous, Riemann integrable, or bounded? The answers to these questions are no in general.

- If a sequence of continuous functions f_k on $[a, b]$ converges uniformly on $[a, b]$ to a function f, it is known that f is continuous, therefore Riemann integrable, and $\int_a^b |f_k(x) - f(x)|\,dx \to 0$ as $k \to \infty$. Furthermore, the space $C[a, b]$ of continuous functions with $\|f\|_{C[a,b]} := \max_{[a,b]} |f(x)|$ as a norm is a complete normed space.

  Is there an analogue in $C[a, b]$ of the Bolzano-Weierstrass Theorem on $\mathbb{R}$ that *any bounded sequence has a convergent subsequence*? What if we know that a sequence in $C[a, b]$ already converges pointwise on a countable and dense subset of $[a, b]$? Is this enough to imply that the sequence would converge pointwise or uniformly? The answer turns out to be no in general, and a positive conclusion would require **equi-continuity** on the sequence.

- $\|f\|_{L[a,b]} := \int_a^b |f(x)|\,dx$ also defines a norm on $C[a, b]$. Does this norm also make $C[a, b]$ a complete normed space? The answer is no, and we will discuss in some detail characterization of elements in the completion of $C[a, b]$ under this and other integral norms.

We will also discuss relations between convergence in integral norms and pointwise or uniform convergence.

The issues above also have their analogues in the context of sequences and series: We know that if $a_{k,l} \to b_l$ as $k \to \infty$ for each l and that if $\sum_{l=1}^{\infty} a_{k,l}$ converges for each k, then $\sum_{l=1}^{\infty} b_l$ does not necessarily converge, and even if it does, it may not equal $\lim_{k\to\infty} \sum_{l=1}^{\infty} a_{k,l}$ when the latter exists, even if $a_{k,l} \to b_l$ converges uniformly in l. The relevant condition needed here is **equi-summability** of the sequence of series. In the general theory of integration, these issues are usually addressed by the so-called *dominated convergence theorem.* The concrete contexts here reveal the central issues in addressing such convergence problems.

To prepare for our discussion, we will first briefly discuss the Riemann-Stieltjes integral at the beginning of chapter 1. We then discuss properties of elements in the completion of the space $C[a, b]$ of continuous functions under integral norms adapted from an approach taken by P. Lax. Here we discuss the essential features of Lebesgue's integral on $\mathbb{R}$ without relying on the measure theory on the so-called σ-algebra of measurable sets. Finally we discuss the notions of absolutely continuous functions and of bounded variation, and their relations to the extension of the Fundamental Theorem of Calculus.

Chapter 2 records the most useful properties related to the convergence of sequences and series of functions. Chapter 3 records the basic properties of power series. A large part of these two chapters is usually done in the first semester.

Chapter 4 discusses the basics of Fourier series. Because there may not be enough time to have a thorough discussion on the properties of elements in the completion of the space $C[a, b]$ of continuous functions under integral norms, Chapter 4 is written in a way that can be studied without a detailed knowledge of the functions in the completion references above.

Chapters 5--7 focus on the differential and integral calculus of functions of several variables. Here my approach is very close to that of Spivak's, but I also try to provide more motivated discussions on the origin of the basic notions involved: curl and divergence of vector fields and differential of forms.

The construction of the course notes was sponsored by an 'Open and Affordable Textbooks Program' (OAT Program) award from the Rutgers University Libraries in spring 2023.

The OAT Program supports textbook affordability at Rutgers by encouraging courses to adopt educational materials that are freely available, available at a low cost (compared to similar courses), or part of the Rutgers University Libraries' electronic collections, and thereby free of charge to Rutgers University students.

Contents

Back Matter

Chapter 1

Select Topics on Integration in One Variable

1.1 A Brief Summary of Riemann-Stieltjes Integral

In the context of probability, if a random variable has a probability density function $p(x)$, then the probability that the variable falls into $[a,b]$ is given by $\int_a^b p(x)\,dx$. Such a random variable can't have a positive probability taking on a single value. A general random variable X has a cumulative distribution function $F(x) := \mathcal{P}(x : X \leq x)$. It is a non-decreasing function of x---we will also use the terminology increasing function interchangeably. The Riemann-Stieltjes integral with respect to an increasing function is a step in describing the probability of a general random variable in terms of integrals.

In the context of functional analysis, there is a need to consider possible convergence of a sequence of linear functional on $C[a,b]$, given in the form of $g \in C[a,b] \mapsto \int_a^b g(x) f_k(x)\,dx$, where $\{f_k\}$ is a sequence of continuous or Riemann integrable functions on $[a,b]$ with certain bounds, say, $\int_a^b |f_k(x)|\,dx \leq M$ for all k. If there is a limit, then it could be given in terms of the Riemann-Stieltjes integral with respect to some increasing function α in the sense that $\int_a^b g(x) f_k(x)\,dx \to \int_a^b g\,d\alpha$ (to be defined below).

Given a *monotone increasing function* α on $[a,b]$ (this will be a standing assumption throughout this chapter), we can mimic the steps in defining the usual Riemann integral of a given bounded function f on $[a,b]$ to define the lower sum

$$L(f, \mathcal{P}, d\alpha) := \sum_{i=1}^{k} m_{I_i}(f) d\alpha(I_i)$$

and upper sum

$$U(f, \mathcal{P}, d\alpha) := \sum_{i=1}^{k} M_{I_i}(f) d\alpha(I_i)$$

of f with respect to a partition $\mathcal{P} = \{a = x_0 < x_1 < \cdots < x_k = b\}$ of $[a,b]$, with $I_i = [x_{i-1}, x_i]$, $d\alpha(I_i) = \alpha(x_i) - \alpha(x_{i-1})$, $m_{I_i}(f) = \inf_{I_i} f$, and $M_{I_i}(f) = \sum_{I_i} f$.

Definition 1.1.1

Suppose that f is a bounded function defined on the interval $[a, b]$. Then its upper integral, and respectively lower integral, on $[a, b]$ with respect to α is defined as $\inf_{\mathcal{P}} U(f, \mathcal{P}, \alpha)$, and respectively $\sup_{\mathcal{P}} L(f, \mathcal{P}, \alpha)$, where $\mathcal{P}$ runs over all partitions of $\mathcal{P}$.

The upper integral is denoted as $\overline{\int}_a^b f\,d\alpha$, while the lower integral is denoted as $\underline{\int}_a^b f\,d\alpha$.

When $\alpha(x) = x$, we write $U(f, \mathcal{P})$ for $U(f, \mathcal{P}, dx)$, and $L(f, \mathcal{P})$ for $L(f, \mathcal{P}, dx)$.

Exercise 1.1.2 Let $\mathcal{C}$ denote the standard tertiary Cantor set on $[0, 1]$ and $\chi_{\mathcal{C}}$ denote its characteristic function which takes value 1 on $\mathcal{C}$ and 0 elsewhere. Let $\mathcal{P}$ be a partition of $[0, 1]$ into intervals of equal length 3^{-k} for some $k \in \mathbb{N}$. Find $U(\chi_{\mathcal{C}}, \mathcal{P})$ and $L(\chi_{\mathcal{C}}, \mathcal{P})$. Is there a positive lower bound of $U(\chi_{\mathcal{C}}, \mathcal{P})$ independent of $\mathcal{P}$?

Proposition 1.1.3 Basic Property of Upper and Lower Integral of a Bounded Real-valued Function.

Suppose that f is a bounded function defined on the interval $[a, b]$. Then

$$\underline{\int}_a^b f\,d\alpha \leq \overline{\int}_a^b f\,d\alpha.$$

Proof. Let $\mathcal{P}_1, \mathcal{P}_2$ be two arbitrary partitions of $[a, b]$, and $\mathcal{P}^*$ be a refinement of both $\mathcal{P}_1$ and $\mathcal{P}_2$. Then

$$L(f, \mathcal{P}_1, d\alpha) \leq L(f, \mathcal{P}^*, d\alpha) \leq U(f, \mathcal{P}^*, d\alpha) \leq U(f, \mathcal{P}_2, d\alpha).$$

As a result,

$$L(f, \mathcal{P}_1, d\alpha) \leq \overline{\int}_a^b f\,d\alpha) = \inf_{\mathcal{P}_2} U(f, \mathcal{P}_2, d\alpha),$$

and

$$\underline{\int}_a^b f\,d\alpha = \sup_{\mathcal{P}_1} L(f, \mathcal{P}_1, d\alpha) \leq \overline{\int}_a^b f\,d\alpha.$$

■

Definition 1.1.4

A bounded real-valued function f defined on an interval $[a, b]$ is called **Riemann-Stieltjes integrable** with respect to $d\alpha$, if

$$\underline{\int}_a^b f\,d\alpha = \overline{\int}_a^b f\,d\alpha.$$

In such a case we write $f \in \mathcal{R}(\alpha)$ on $[a, b]$, and use $\int_a^b f\,d\alpha$ for $\underline{\int}_a^b f\,d\alpha$.

Theorem 1.1.5 Riemann-Stieltjes Integrability Criterion.

A bounded real-valued function f defined on the interval $[a, b]$ is Riemann-Stieltjes integrable iff for any $\epsilon > 0$, there exists a partition $\mathcal{P}$ of $[a, b]$ such that

$$U(f, \mathcal{P}, d\alpha) - L(f, \mathcal{P}, d\alpha) = \sum_{i=1}^{k} (M_{I_i}(f) - m_{I_i}(f))\, d\alpha(I_i) < \epsilon. \tag{1.1.1}$$

Exercise 1.1.6 Define

$$H(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \text{ and } G(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Determine $\underline{\int}_{-1}^{1} H(x)\, dH(x)$, $\overline{\int}_{-1}^{1} H(x)\, dH(x)$, $\underline{\int}_{-1}^{1} G(x)\, dH(x)$, $\overline{\int}_{-1}^{1} G(x)\, dH(x)$, and find out if either H or G is in $\mathcal{R}(dH)$ over $[-1, 1]$.

Theorem 1.1.7

If f is continuous on $[a, b]$, then $f \in \mathcal{R}(\alpha)$ on $[a, b]$.

Proof. Note that in order to make

$$U(f, \mathcal{P}, d\alpha) - L(f, \mathcal{P}, d\alpha) = \sum_{i=1}^{k} (M_{I_i}(f) - m_{I_i}(f))\, d\alpha(I_i) < \epsilon,$$

to establish (1.1.1), we can use the uniform continuity of f on $[a, b]$ to find a partition $\mathcal{P}$ of $[a, b]$ such that each $M_{I_i}(f) - m_{I_i}(f) < \epsilon/(\alpha(b) - \alpha(a))$. It then follows that $U(f, \mathcal{P}, d\alpha) - L(f, \mathcal{P}, d\alpha) < \epsilon$. ■

Theorem 1.1.8

If f is monotone on $[a, b]$ and α (monotone) continuous on $[a, b]$, then $f \in \mathcal{R}(\alpha)$ on $[a, b]$.

Proof. Here we make use of the uniform continuity of α on $[a, b]$ to find a partition $\mathcal{P}$ of $[a, b]$ such that each $\alpha(x_i) - \alpha(x_{i-1}) < \epsilon/(f(b) - f(a))$. It then follows that $U(f, \mathcal{P}, d\alpha) - L(f, \mathcal{P}, d\alpha) < \epsilon$. ■

Theorem 1.1.9

If f is bounded on $[a, b]$ and has only finitely many points of discontinuity, and α is continuous at every point at which f is discontinuous, then $f \in \mathcal{R}(\alpha)$ on $[a, b]$.

Proof. Let $M > 0$ be such that $|f(x)| \leq M$ for all $x \in [a, b]$ and $p_1 < p_2 < \cdots < p_k$ be all the points of discontinuity of f in $[a, b]$ (Here we are assuming $p_1 > a$ and $p_k < b$; the proof easily adapts to the remaining cases). Using the continuity of α at these points, we can find $l_i < p_i < r_i$ such that $r_i < l_{i+1}$ and

$$\sum_{i=1}^{k} \left(M_{[l_i,r_i]}(f) - m_{[l_i,r_i]}(f)\right)(\alpha(r_i) - \alpha(l_i)) < \epsilon/2.$$

Using the uniform continuity of f on $[a,b] \setminus \cup_{i=1}^{k}(l_i, r_i)$, we can find a partition $\mathcal{P}$ on this finite union of closed intervals such that

$$\sum_{j} (M_j(f) - m_j(f))(\alpha(x_j) - \alpha(x_j) < \epsilon/2.$$

Adjoining $\mathcal{P}$ with $\cup_{i=1}^{k}[l_i, r_i]$ forms a partition of $[a,b]$ for which (1.1.1) holds. ■

Theorem 1.1.10 Properties of Riemann-Stieltjes Integral.

Suppose that $f_m1, f_2 \in \mathcal{R}(\alpha)$ on $[a,b]$.

(a) $f_1 + f_2 \in \mathcal{R}(\alpha)$ and $cf_1 \in \mathcal{R}(\alpha)$ on $[a,b]$ for any scalar c, and

$$\int_a^b (f_1 + f_2)\,d\alpha = \int_a^b f_1 d\alpha + \int_a^b f_2\,d\alpha$$
$$\int_a^b cf_1 d\alpha = c\int_a^b f_1 d\alpha.$$

(b) $f_1 f_2 \in \mathcal{R}(\alpha)$.

(c) $|f_1| \in \mathcal{R}(\alpha)$ and $\left|\int_a^b f_1 d\alpha\right| \leq \int_a^b |f_1| d\alpha$.

(d) If $f_1(x) \leq f_2(x)$ on $[a,b]$, then

$$\int_a^b f_1 d\alpha \leq \int_a^b f_2 d\alpha.$$

(e) For any $c, a < c < b$, then $f_1 \in \mathcal{R}(\alpha)$ on $[a,c]$ and on $[c,b]$, and

$$\int_a^c f_1 d\alpha + \int_c^b f_1 d\alpha = \int_a^b f_1 d\alpha.$$

(f) If $|f_1(x)| \leq M$ on $[a,b]$, then

$$\left|\int_a^b f_1 d\alpha\right| \leq M[\alpha(b) - \alpha(b)].$$

(g) If β is another monotone increasing function on $[a,b]$ and $f_1 \in \mathcal{R}(\beta)$ on $[a,b]$, then $f_1 \in cR(\alpha + \beta)$ on $[a,b]$ and

$$\int_a^b f_1 d(\alpha + \beta) = \int_a^b f_1 d\alpha + \int_a^b f_1 d\beta.$$

Theorem 1.1.11

Assume that f is bounded on $[a,b]$ and that $\alpha \in C[a,b]$ and $\alpha' \in \mathcal{R}$ on $[a,b]$ (namely, α' is Riemann integrable on $[a,b]$). Then $f \in \mathcal{R}(\alpha)$ on $[a,b]$ iff $f\alpha' \in \mathcal{R}$ on $[a,b]$, and

$$\int_a^b f\,d\alpha = \int_a^b f\alpha'\,dx. \tag{1.1.2}$$

Proof. We only outline the main ingredients. Let $M = \sup_{[a,b]} |f(x)|$. For any given $\epsilon > 0$, first use $\alpha' \in \mathcal{R}$ on $[a,b]$ and (1.1.1) to find a partition $\mathcal{P} = \{a = x_0 < x_1 < \cdots < x_k = b\}$ such that

$$U(\alpha', \mathcal{P}, dx) - L(\alpha', \mathcal{P}, dx) = \sum_{i=1}^{k} \left(M_{I_i}(\alpha') - m_{I_i}(\alpha')\right)(x_i - x_{i-1}) < \epsilon. \tag{1.1.3}$$

Use this to prove that for any choice $s_i \in [x_{i-1}, x_i]$

$$\left| \sum_{i=1}^{k} f(s_i)(\alpha(x_i) - \alpha(x_{i-1})) - \sum_{i=1}^{k} f(s_i)\alpha'(s_i)(x_i - x_{i-1}) \right| \le M\epsilon. \tag{1.1.4}$$

It follows from

$$\sum_{i=1}^{k} f(s_i)(\alpha(x_i) - \alpha(x_{i-1})) \le \sum_{i=1}^{k} f(s_i)\alpha'(s_i)(x_i - x_{i-1}) + M\epsilon \le U(f\alpha', \mathcal{P}, dx) + M\epsilon$$

that

$$U(f, \mathcal{P}, d\alpha) \le U(f\alpha', \mathcal{P}, dx) + M\epsilon.$$

Reversing the roles of $U(f, \mathcal{P}, d\alpha)$ and $U(f\alpha', \mathcal{P}, dx)$ leads to

$$|U(f, \mathcal{P}, d\alpha) - U(f\alpha', \mathcal{P}, dx)| \le M\epsilon. \tag{1.1.5}$$

By definition there exist partitions $\mathcal{P}_1$ and $\mathcal{P}_2$ such that

$$\overline{\int_a^b} f\,d\alpha \le U(f, \mathcal{P}_1, d\alpha) < \overline{\int_a^b} f\,d\alpha + \epsilon \tag{1.1.6}$$

and

$$\overline{\int_a^b} f\alpha'\,dx \le U(f\alpha', \mathcal{P}_2, dx) < \overline{\int_a^b} f\alpha'\,dx + \epsilon. \tag{1.1.7}$$

Let $\mathcal{P}^*$ be a common refinement of $\mathcal{P}, \mathcal{P}_1, \mathcal{P}_2$, then (1.1.3)--(1.1.7) continue to hold with $\mathcal{P}^*$ replacing $\mathcal{P}, \mathcal{P}_1, \mathcal{P}_2$, respectively. It now follows that

$$\begin{aligned} \overline{\int_a^b} f\alpha'\,dx &\le U(f\alpha', \mathcal{P}^*, dx) \le U(f, \mathcal{P}^*, d\alpha) + M\epsilon \\ &\le \overline{\int_a^b} f\,d\alpha + (M+1)\epsilon. \end{aligned}$$

Reversing the roles of $\overline{\int_a^b} f\alpha'\,dx$ and $\overline{\int_a^b} f\,d\alpha$ would lead to

$$\overline{\int_a^b} f\,d\alpha \le \overline{\int_a^b} f\alpha'\,dx + (M+1)\epsilon.$$

Since $\epsilon > 0$ is arbitrary, we conclude that

$$\overline{\int_a^b} f\, d\alpha = \overline{\int_a^b} f\alpha'\, dx. \tag{1.1.8}$$

Similarly,

$$\underline{\int_a^b} f\, d\alpha = \underline{\int_a^b} f\alpha'\, dx. \tag{1.1.9}$$

Thus $\overline{\int_a^b} f\, d\alpha = \underline{\int_a^b} f\, d\alpha$ iff $\overline{\int_a^b} f\alpha'\, dx = \underline{\int_a^b} f\alpha'\, dx$. ■

Corollary 1.1.12

Assume that $\alpha \in C[a,b]$ and $\alpha' \in \mathcal{R}$ on $[a,b]$. Then for any $t \in [a,b]$

$$\int_a^t \alpha'(x)\, dx = \alpha(t) - \alpha(a). \tag{1.1.10}$$

Remark 1.1.13

If α has a finite number of jump discontinuity, then the right hand side of (1.1.2) and (1.1.10) need to account for contributions from these points. The assumption $\alpha' \in \mathcal{R}$ on $[a,b]$ is needed. Using a construction similar to that of a Cantor set, Volterra constructed a function (not monotone though) which had derivative everywhere in $[0,1]$ but its derivative is not Riemann integrable. Cantor's function is monotone increasing in $[0,1]$, equals a constant in any of the middle third interval removed, therefore has derivative equal 0 at any point of those intervals. Since Cantor's set has measure 0, Cantor's function has derivative equal 0 almost everywhere in $[0,1]$. If we can define an integral for such a function, its integral on any subinterval of $[0,1]$ must be 0, thus the above equality relation can't hold in such a case.

Remark 1.1.14

Suppose that $f \in \mathcal{R}(\alpha)$ on $[a,b]$ and $a < c < b$. Then $\int_a^c 1\, d\alpha = \alpha(c) - \alpha(a)$, $\lim_{k\to\infty} \int_a^{c+\frac{1}{k}} 1\, d\alpha = \alpha(c+0) - \alpha(a)$. Since $[a,c] = \cap_k [a, c+\frac{1}{k}]$, or one could think of $\int_a^c 1\, d\alpha$ as $\int \chi_{[a,c]}(x)\, d\alpha$, and since $\lim_{k\to\infty} \chi_{[a,c+\frac{1}{k}]}(x) = \chi_{[a,c]}(x)$, one expects $\lim_{k\to\infty} \int_a^{c+\frac{1}{k}} 1\, d\alpha = \int_a^c 1\, d\alpha$. This would require $\alpha(c+0) = \alpha(c)$. For this reason, one often chooses to work with an α which is right-continuous.

Since $\lim_{k\to\infty} \int_a^{c-\frac{1}{k}} 1\, d\alpha = \alpha(c-0) - \alpha(a)$ and $[a,c) = \cup_k [a, c-\frac{1}{k}]$, it is reasonable to treat $\lim_{k\to\infty} \int_a^{c-\frac{1}{k}} 1\, d\alpha$ as $\int_{[a,c)} 1\, d\alpha$ and $\lim_{k\to\infty} \int_a^{c+\frac{1}{k}} 1\, d\alpha$ as $\int_{[a,c]} 1\, d\alpha$; namely, $\int_{[a,c)} 1\, d\alpha$ and $\int_a^c d\alpha = \int_{[a,c]} 1\, d\alpha$ carry different meaning--- the notation of integration over a set is more precise than that of integration over a lower and upper limit.

1.2 The Completion of the Space of Continuous Functions under Integral Norms

We sketch below some main properties of the completion of the space of continuous functions under integral norms. The main line of approach is adapted from the following article by P. Lax, *Rethinking the Lebesgue Integral*, The American Mathematical Monthly, Dec., 2009, Vol. 116, No. 10 (Dec., 2009), pp. 863-881. Although the discussion can be done with respect to the Riemann-Stieltjes integral, we limit our discussion to the standard Riemann integral.

The main feature of this discussion is to establish the important property that the class of integrable functions is closed under reasonable sequential limit while avoiding the discussion of the so called σ-algebra of measurable sets. Any set can be identified by its characteristic function, and the union and intersection of a finite number or countably many sets can also be represented in terms of algebraic sum or limits of the characteristic functions of the relevant sets, so when a sequence of characteristic functions has a limit in the integral sense, it gives information about the limit of the corresponding sets.

For any $1 \le p < \infty$, let $L^p[a,b]$ denote the completion of $C[a,b]$ under the norm $\|f\|_{L^p[a,b]} := \left(\int_a^b |f(x)|^p \, dx\right)^{1/p}$. From the abstract completion process, each element $\phi \in L^p[a,b]$ is an equivalence class of Cauchy sequences $\{c_k\}$ in $C[a,b]$ under the $L^p[a,b]$-norm. We would like to address the following questions

a. Does each equivalence class of Cauchy sequences $\{c_k\}$ in $C[a,b]$ under the $L^p[a,b]$-norm associate to a pointwise-defined function on $[a,b]$?

b. Does each equivalence class of Cauchy sequences $\{c_k\}$ in $C[a,b]$ under the $L^p[a,b]$-norm associate to a well-defined integral? What usual properties of integrals are preserved?

c. Establish criteria for other limiting process of functions in $C[a,b]$ or $L^p[a,b]$, e.g., pointwise limit, that result in a limit in $L^p[a,b]$.

An elementary observation. Suppose that $\{c_k\}$ is a Cauchy sequence in $C[a,b]$ under the $L^p[a,b]$-norm. Then for $[a',b'] \subset [a,b]$, $\int_{a'}^{b'} c_k(x)\,dx$ is a Cauchy sequence in $\mathbb{R}$; and for any $1 \le q \le p$, $\|c_k\|_{L^q[c,d]}$ is a Cauchy sequence in $\mathbb{R}$. Furthermore, $\lim_{k\to\infty} \int_{a'}^{b'} c_k(x)\,dx$ and $\lim_{k\to\infty} \|c_k\|_{L^q[c,d]}$ remain the same if $\{c_k\}$ is replaced by an equivalent Cauchy sequence.

These properties are simple consequences of the triangle inequalities:

$$\begin{aligned}
|\int_{a'}^{b'} c_k(x)\,dx - \int_{a'}^{b'} c_l(x)\,dx| &\le \int_{a'}^{b'} |c_k(x) - c_l(x)|\,dx \\
&\le \left(\int_{a'}^{b'} |c_k(x) - c_l(x)|^p\,dx\right)^{1/p} (b'-a')^{1-1/p}, \\
\big|\|c_k\|_{L^q[a',b']} - \|c_l\|_{L^q[a',b']}\big| &\le \|c_k - c_l\|_{L^q[a',b']} \\
&\le \|c_k - c_l\|_{L^p[a',b']}(b'-a')^{1/q-1/p}.
\end{aligned}$$

As a result, if we denote by ϕ the equivalence class of $\{c_k\}$, it makes sense to define

$$\int_{a'}^{b'} \phi = \lim_{k\to\infty} \int_{a'}^{b'} c_k(x)\,dx, \quad \|\phi\|_{L^q[a',b']} = \lim_{k\to\infty} \|c_k\|_{L^q[a',b']}.$$

Any $c \in C[a,b]$ defines a Cauchy sequence $\{c_k = c\}$ in $L^p[a,b]$ and $\|c\|_{L^q[c,d]}$ as defined through the limiting process above equals the definition through Riemann's integral.

It also follows routinely that if $\{[c_i,d_i]\}$ is a finite collection of disjoint subintervals of $[a,b]$, then

$$\int_{\cup_i [c_i,d_i]} \phi = \sum_i \int_{[c_i,d_i]} \phi$$

is well defined; the intervals $\{[c_i, d_i]\}$ can also be replaced by open or half-open intervals. Furthermore, if $\phi, \psi \in L^p[a,b]$, then for any scalars α, β,

$$\int_{[a,b]} (\alpha\phi + \beta\psi) = \alpha \int_{[a,b]} \phi + \beta \int_{[a,b]} \psi.$$

Any open set G of $\mathbb{R}$ is a union of at most countably many disjoint open intervals (c_i, d_i). For any $c \in C[a,b]$ and any open subset G of $[a,b]$, the integrals $\int_G c(x)\,dx$ and $\int_G |c(x)|^p\,dx$ have clearly defined meanings as the (absolutely convergent) sum of the integrals on each constitutive open interval, with $\int_G 1\,dx = |G|$, the length of G .

Definition 1.2.1 Uniform Absolute Continuity of A Sequence of Integrals.

The L^p-integrals of a sequence of functions $\{c_k\}$ in $C[a,b]$ are said to be **uniformly absolutely continuous** on $[a,b]$ if for any $\epsilon > 0$, there exists a $\delta > 0$ such that for any open subset G of $[a,b]$ with its length $|G| < \delta$,

$$\int_G |c_k(x)|^p\,dx < \epsilon \text{ for all } k.$$

Proposition 1.2.2 Uniform Absolute Continuity of the Integrals of a Cauchy Sequence in $C[a,b]$.

The L^p-integrals of any Cauchy sequence $\{c_k\}$ in $C[a,b]$ under the $L^p[a,b]$-norm are uniformly absolutely continuous on $[a,b]$.

Proof. For the given $\epsilon > 0$, there exists some N such that for all $k, l \geq N$, $\|c_k - c_l\|_{L^p[a,b]} < \epsilon^{1/p}/2$. Since $c_i \in C[a,b]$ for $i = 1, \cdots, N$, there exists a common $M > 0$ such that $|c_k(x)| \leq M$ for all $x \in [a,b]$ and $1 \leq k \leq N$, therefore, there exists $\delta > 0$ such that for any open subset G of $[a,b]$ with its length $|G| < \delta$, we have

$$\int_G |c_k(x)|^p\,dx < \epsilon/2^p \text{ for } 1 \leq k \leq N.$$

Now for $l > N$, the triangle inequality continues to hold for integrals on G, which implies that

$$\|c_l\|_{L^p(G)} \leq \|c_l - c_N\|_{L^p(G)} + \|c_N\|_{L^p(G)} \leq \epsilon^{1/p}/2 + \epsilon^{1/p}/2.$$

from which our conclusion follows. ■

Exercise 1.2.3 In the context of Proposition 1.2.2, suppose that the open set G is decomposed as the union of at most countably many disjoint open intervals: $G = \cup_l I_l$, and set $a_{k,l} = \int_{I_l} |c_k(x)|^p\,dx$. Prove that

$$\lim_{k\to\infty} \sum_l a_{k,l} = \sum_l \lim_{k\to\infty} a_{k,l}.$$

This allows to define $\int_G |\phi(x)|^p\,dx$ as $\sum_l \int_{I_l} |\phi(x)|^p\,dx$, given by $\lim_{k\to\infty} \int_G |c_k(x)|^p\,dx$.

As a consequence, prove the **absolute continuity of the integral** of $|\phi|^p$, namely, for any $\phi \in L^p[a,b]$ and for any $\epsilon > 0$, there exists a $\delta > 0$ such that for any open subset G of $[a,b]$ with its length $|G| < \delta$,

$$\int_G |\phi(x)|^p\,dx < \epsilon.$$

Exercise 1.2.4 Construct $b_{k,l} \geq 0$ for k, l such that $\lim_{k\to\infty} b_{k,l}$ exists for each l (the convergence can even be uniform in l) and

$$\lim_{k\to\infty} \sum_l b_{k,l} \neq \sum_l \lim_{k\to\infty} b_{k,l}.$$

Exercise 1.2.5 Construct a sequence of non-negative functions $\{c_k(x)\}$ in $C[0,1]$ such that $\lim_{k\to\infty} c_k(x) = 0$ for each $x \in [0,1]$ but $\lim_{k\to\infty} \int_0^1 c_k(x)\,dx > 0$. Verify that the integrals of this sequence fail to be uniformly absolutely continuous on $[0,1]$.

Definition 1.2.6 Negligible Set.

A point set is called negligible if it can be covered by an open set of arbitrarily small volume, namely, for any $\epsilon > 0$, there exists an open cover G such that $|G| < \epsilon$.

The notion of negligible set here is a special case in the context of Lebesgue measure on $\mathbb{R}$ of the notion of a set of measure 0. Note that the union of at most countable number of negligible sets is still negligible.

When two functions are equal except on a negligible set, we use the customary language that they are equal *a.e.*.

Definition 1.2.7 Realization of a Cauchy Sequence in $L^p[a,b]$.

A function $\phi(x)$ defined *a.e.* on $[a,b]$ is said to be a realization of $\phi \in L^p[a,b]$ if there exists a Cauchy sequence $\{c_k\}$ of continuous functions in $[a,b]$ in the equivalence class of ϕ which converges a.e to $\phi(x)$:

$$\lim_{k\to\infty} c_k(x) \to \phi(x) \text{ } a.e..$$

Note that if I is any interval of finite length (either open or closed) in $[a,b]$, then for any k, one can find a continuous function $c_k(x)$ with support in the interior of I, such that $\int |\chi_I(x) - c_k(x)|\,dx < 1/k$. This shows that the characteristic function $\chi_I(x)$ of I is a realization of a function in $L^p[a,b]$, namely, it is a function in $L^p[a,b]$. The same applies to the linear combination of a finite number of such characteristic functions.

Since the Riemann sum of a function is the integral of such a linear combination of a finite number of such characteristic functions. and any Riemann integrable function on $[a,b]$ can be approximated in $L[a,b]$ by its Riemann sums, we conclude that any Riemann integrable function is a realization of some function in $L^p[a.b]$.

A similar argument shows that the characteristic function of any open set of $\mathbb{R}$ with finite length is a realization of some function in $L^p[a.b]$.

Proposition 1.2.8 Realization of Elements of $L^p[a,b]$.

(i) *Any Cauchy sequence $\{c_k\}$ in $C[a,b]$ under the $L^p[a,b]$-norm has a subsequence which converges pointwise a.e. on $[a,b]$. Furthermore, for any $\epsilon > 0$, there exists an open set G with $|G| < \epsilon$ such that $\{c_k\}$ converges uniformly over G^c.*

(ii) *If $\{c_k\}$ and $\{\tilde{c}_k\}$ are two equivalent Cauchy sequences in $C[a,b]$ under the $L^p[a,b]$-norm such that both $\lim_{k\to\infty} c_k(x)$ and $\lim_{k\to\infty} \tilde{c}_k(x)$ exist*

a.e., then $\lim_{k\to\infty} c_k(x) = \lim_{k\to\infty} \tilde{c}_k(x)$ *a.e..*

As a result, any ϕ *in* $L^p[a,b]$*, namely, an equivalence class of Cauchy sequences in* $C[a,b]$ *under the* $L^p[a,b]$*-norm, has an a.e. pointwise defined realization defined on* $[a,b]$*. We will denote such a realization by* $\phi(x)$*. In particular, if* ϕ *in* $L^p[a,b]$ *is such that* $\|\phi\|_{L^p[a,b]} = 0$*, then* $\phi(x) = 0$ *a.e. on* $[a,b]$*.*

(iii) If $\{c_k\}$ *and* $\{\tilde{c}_k\}$ *are two Cauchy sequences in* $C[a,b]$ *under the* $L^p[a,b]$*-norm such that* $\lim_{k\to\infty} c_k(x) = \lim_{k\to\infty} \tilde{c}_k(x)$ *a.e., then* $\{c_k\}$ *and* $\{\tilde{c}_k\}$ *are equivalent Cauchy sequences in* $L^p[a,b]$*.*

Proof. Any Cauchy sequence $\{c_k\}$ in $C[a,b]$ under the $L^p[a,b]$-norm has a subsequence, still denoted as $\{c_k\}$, such that

$$\|c_k - c_{k+1}\|_{L^p[a,b]} \le \epsilon_k^2,$$

where $\epsilon_k > 0$ are such that $\sum_k \epsilon_k$ converges.

Define

$$D_k = \{x \in [a,b] : |c_k(x) - c_{k+1}(x)| > \epsilon_k\}.$$

Then D_k is open and it follows from

$$\epsilon_k^p |D_k| \le \int_{[a,b]} |c_k(x) - c_{k+1}(x)|^p \, dx \le \epsilon_k^{2p}$$

that $|D_k| \le \epsilon_k^p$.

Note that $\cup_{k=N}^\infty D_k$ is open and

$$|\cup_{k=N}^\infty D_k| \le \sum_{k=N}^\infty |D_k| \le \sum_{k=N}^\infty \epsilon_k^p \to 0,$$

as $N \to \infty$. Here, the D_k's may not be disjoint, but this subadditivity property used above holds. Therefore, the set E^1 defined by

$$E = \cap_{N=1}^\infty \cup_{k=N}^\infty D_k$$

is negligible, and for any $x \in E^c$, there exists some N_x such that $x \in \cap_{k=N_x}^\infty D_k^c$, which makes $\{c_k(x)\}$ a Cauchy sequence in $\mathbb{R}$, therefore $\lim_{k\to\infty} c_k(x)$ exists. Furthermore, for any $\epsilon > 0$, there exists some N_ϵ such that $|\cup_{k=N_\epsilon}^\infty D_k| < \epsilon$. On the complement of $\cup_{k=N_\epsilon}^\infty D_k$, $|c_k(x) - c_{k+1}(x)| \le \epsilon_k$ for all $k \ge N_\epsilon$, implying that $\{c_k(x)\}$ converges uniformly on this set.

For (ii), we can use the equivalent Cauchy sequences $\{c_k\}$ and $\{\tilde{c}_k\}$ to construct a new Cauchy sequence, say, with c_k as the $(2k-1)$-th term and $\{\tilde{c}_k\}$ as the $(2k)$-th term, then appeal to the proof of (i), making sure that in selecting the subsequence, infinitely many terms from both sequences are selected. This subsequence, selected from two *a.e.* convergent sequences, converges *a.e.*, which shows that $\lim_{k\to\infty} c_k(x) = \lim_{k\to\infty} \tilde{c}_k(x)$ *a.e.*.

For (iii), define $f_k(x) = c_k(x) - \tilde{c}_k(x)$, then $\{f_k(x)\}$ is a Cauchy sequence in $C[a,b]$ under the $L^p[a,b]$-norm such that $\lim_{k\to\infty} f_k(x) = 0$ *a.e.*. We now show that $\{f_k(x)\} \to 0$ in $L^p[a,b]$. We argue by contradiction: suppose not, then there exists some $\sigma > 0$ and a subsequence of $\{f_k(x)\}$, still denoted as $\{f_k(x)\}$, such that $\int_a^b |f_k(x)|^p \, dx \ge \sigma$ for all k.

For any $\epsilon > 0$, first apply Proposition 1.2.2 to $\{f_k(x)\}$ to find $\delta > 0$ such that for any open set G in $[a,b]$ with $|G| < \delta$, we have $\int_G |f_k(x)|^p \, dx < \epsilon$. Next apply the proof for (i) to $\{f_k(x)\}$ to find a subsequence of $\{f_k(x)\}$, still denoted as $\{f_k(x)\}$, a limit function $\phi(x)$ *a.e.* defined, and an open set G in $[a,b]$ with $|G| < \delta$ such that $f_k(x) \to \phi(x)$ *a.e.* and uniformly in G^c. Since it is assumed that $f_k(x) \to 0$ *a.e.*, we must have $\phi(x) = 0$ *a.e.*. Thus there exists N such that $|f_k(x)|^p \le \epsilon$ for all $x \in G^c$ and $k \ge N$.

We would like to estimate $\int_{[a,b]} |f_k(x)|^p\,dx$ by $\int_G |f_k(x)|^p\,dx + \int_{G^c} |f_k(x)|^p\,dx$. But in the Riemann integral setting, integration over an arbitrary closed set is not defined. We complete the argument in the following way. For any $k > N$, since $|f_k(x)|^p \in \mathcal{R}[a,b]$, we can find a partition $a = a_0 < a_1 < \cdots < a_{N_k} = b$ such that

$$\left| \int_a^b |f_k(x)|^p\,dx - \sum_{i=1}^{N_k} m_i(|f_k|^p)(a_i - a_{i-1}) \right| < \epsilon.$$

For any interval $[a_{i-1}, a_i]$ that has non-empty intersection with G^c, we see that $m_i(|f_k|^p) = \inf_{[a_{i-1},a_i]} |f_k|^p \le \epsilon$; the remaining subintervals $[a_{i-1}, a_i]$ are contained in G, and the lower Riemann sum over such intervals $\le \int_G |f_k(x)|^p\,dx \le \epsilon$ for any $k \ge N$. It then follows that for any $k \ge N$,

$$\int_a^b |f_k(x)|^p\,dx \le (b-a)\epsilon + 2\epsilon.$$

Choose $\epsilon > 0$ at the beginning such that $(b-a)\epsilon + 2\epsilon < \sigma$, this then shows a contradiction with our assumption that $\int_a^b |f_k(x)|^p\,dx \ge \sigma$ for all k. Thus we have shown that $\int_a^b |f_k(x)|^p\,dx \to 0$ as $k \to \infty$. ■

Remark 1.2.9

As a consequence of Proposition 1.2.8, for any realization $\phi(x)$ of some $\phi \in L^p[a,b]$ and any $\epsilon > 0$, there exists a closed set F of $[a,b]$ with $|F^c| < \epsilon$ such that the restriction of $\phi(x)$ on F is continuous. Note that this is not saying that $\phi(x)$ is continuous at every point of F as a function on $[a,b]$. Furthermore, for any $\phi \in L^p[a,b]$ and any $\epsilon, \epsilon' > 0$, one can find a continuous approximation $c \in C[a,b]$ in the sense that exists an open set G with $|G| < \epsilon'$ such that

$$\|\phi - c\|_{L^p[a,b]} < \epsilon, \quad \text{and } |\phi(x) - c(x)| < \epsilon \text{ for all } x \in G^c.$$

Here $\|\phi - c\|_{L^p[a,b]}$ is defined through the limiting process instead of the Riemann integral for a general $\phi \in L^p[a,b]$.

Exercise 1.2.10 Let $f \in L^1[a,b]$. Extend f to be 0 outside of $[a,b]$. Prove that $\int_a^b |f(x+h) - f(x)|\,dx \to 0$ as $|h| \to 0$.

Exercise 1.2.11 Let $f \in L^1[a,b]$. Prove that $F(t) := \int_a^t f(x)\,dx$ is a continuous function of $x \in [a,b]$.

Exercise 1.2.12 Let $f \in L^1[a,b]$. Extend f to be 0 outside of $[a,b]$ and define $f_h(x) = h^{-1}\int_x^{x+h} f(y)\,dy$ for $h > 0$ small. Prove that $\int_a^b |f_h(x) - f(x)|\,dx \to 0$ as $h \to 0$.

[1] E is the set of points that belong to infinitely many D_k's, and E^c is then the set of points that belong to at most a finite number of D_k's.

Corollary 1.2.13 $L^p[a,b]$ is closed under taking absolute values and maximums.

If $\{c_k\}$ in $C[a,b]$ is a Cauchy sequence in $L^p[a,b]$-norm, representing ϕ in $L^p[a,b]$, then $\{|c_k|\}$ in $C[a,b]$ is a Cauchy sequence in $L^p[a,b]$-norm whose realization is given by $|\phi(x)|$.

If $\phi, \psi \in L^p[a,b]$, then $\max\{\phi,\psi\} := (\phi+\psi+|\phi-\psi|)/2 \in L^p[a,b]$ and $\min\{\phi,\psi\} := (\phi+\psi-|\phi-\psi|)/2 \in L^p[a,b]$.

Proposition 1.2.14 Rapidly Convergent Sequence in $L^p[a,b]$.

Let $\{\phi_k\}$ be a sequence of elements in $L^p[a,b]$ converging rapidly in the sense

$$\|\phi_k - \phi_{k+1}\|_{L^p[a,b]} < \epsilon_k^2$$

where $\sum_k \epsilon_k$ converges. Then there exists a unique $\phi \in L^p[a,b]$ such that $\|\phi_k - \phi\|_{L^p[a,b]} \to 0$ and $\phi_k(x) \to \phi(x)$ a.e. as $k \to \infty$.

Proof. For each ϕ_k there exists some $c_k \in C[a,b]$ and an open set G_k with $|G_k| < \epsilon_k$ such that

$$\|\phi_k - c_k\|_{L^p[a,b]} < \epsilon_k^2, \text{ and } |\phi_k(x) - c_k(x)| < \epsilon_k \text{ for all } x \in G_k^c.$$

Then the set $E = \cap_{l=1}^{\infty} \cup_{k=l}^{\infty} G_k$ is negligible, as $|\cup_{k=l}^{\infty} G_k| \le \sum_{k=l}^{\infty} |G_k| \to 0$ as $l \to \infty$, and for any $x \in E^c$, there exists some l_x such that $x \in \cap_{k=l_x}^{\infty} G_k^c$. This implies that $\{c_k\}$ satisfies

$$\|c_k - c_{k+1}\|_{L^p[a,b]} < 2\epsilon_k^2 + \epsilon_{k+1}^2 \text{ and } |\phi_k(x) - c_k(x)| < \epsilon_k \text{ for } k \ge l_x.$$

Then Proposition 1.2.8 applied to $\{c_k\}$ implies that there exists a unique $\phi \in L^p[a,b]$ such that

$$\|c_k - \phi\|_{L^p[a,b]} \to 0 \text{ and } c_k(x) \to \phi(x) \textit{ a.e.} \text{ as } k \to \infty.$$

We can then conclude our proof by appealing to the triangle inequalities. ■

Definition 1.2.15 Nonnegative Elements of $L^p[a,b]$.

An element $\phi \in L^p[a,b]$ is said to be non-negative if its realization $\phi(x)$ satisfies $\phi(x) \ge 0$ *a.e.* on $[a,b]$.

Proposition 1.2.16 Characterization of Nonnegative Elements of $L^p[a,b]$.

An element $\phi \in L^p[a,b]$ is non-negative iff there exists a Cauchy sequence $\{c_k\}$ representing ϕ such that $c_k^- := \min\{c_k(x), 0\}$ satisfies $\|c_k^-\|_{L^p} \to 0$ as $k \to \infty$.

As a consequence, if $\phi \in L^p[a,b]$ is non-negative, then there exists a Cauchy sequence $\{c_k\}$ in $C[a,b]$ representing ϕ such that $\|c_k^+ - \phi\|_{L^p} \to 0$ as $k \to \infty$ and $\int_{[a,b]} \phi = \lim_{k\to\infty} \int_{[a,b]} c_k^+ \, dx \ge 0.$[2]

Proof. Suppose that $\{c_k\}$ is a sequence in $C[a,b]$, Cauchy in $L^p[a,b]$-norm, representing ϕ such that $\|c_k^-\|_{L^p} \to 0$ as $k \to \infty$. Define $c_k^+ = c_k - c_k^-$, which is identified as $\max\{c_k(x), 0\}$. Then $\{c_k^+\}$ is a sequence in $C[a,b]$, Cauchy in $L^p[a,b]$-norm. Proposition 1.2.8 applied to both $\{c_k^+\}$ and $\{c_k^-\}$ implies that there exists a subsequence, still indexed by k to avoid extra indices, such that $\lim_{k\to\infty} c_k^+(x)$ exists *a.e.*, which is evidently ≥ 0 *a.e.*, and $\lim_{k\to\infty} c_k^-(x) \to 0$ *a.e.* (refer to (ii) of Proposition 1.2.8). Therefore

$$\lim_{k\to\infty} c_k(x) = \lim_{k\to\infty} c_k^+(x) + \lim_{k\to\infty} c_k^-(x) \geq 0 \; a.e.$$

Conversely, suppose that $\phi(x) \geq 0$ *a.e.* and $\{c_k\}$ is a sequence in $C[a,b]$, Cauchy under the $L^p[a,b]$ norm representing ϕ and $c_k(x) \to \phi(x)$ *a.e.*. Then we must have $\lim_{k\to\infty} c_k^-(x) \to 0$ *a.e.*. Noting that

$$|c_k^-(x) - c_l^-(x)| \leq |c_k(x) - c_l(x)|$$

we see that $\{c_k^-\}$ is also a sequence in $C[a,b]$, Cauchy under the $L^p[a,b]$ norm. Then the proof of of Proposition 1.2.8 shows that $c_k^- \to 0$ in $L^p[a,b]$ norm, and $\|c_k^+ - \phi\|_{L^p[a,b]} \leq \|c_k - \phi\|_{L^p[a,b]} + \|c_k^-\|_{L^p[a,b]} \to 0$ as $k \to \infty$. ■

Since a function in $L^p[a,b]$ may not be bounded, we may no longer conclude that $f(x)g(x) \in L[a,b]$ whenever $f(x), g(x) \in L[a,b]$. However we have the following

Proposition 1.2.17

(a). Suppose that $f(x) \in L^p[a,b]$ and $g(x) \in L^{p'}[a,b]$ for some $p, p' \geq 1$ such that $1/p + 1/p' = 1$. Then $f(x)g(x) \in L[a,b]$ and

$$\int_a^b |f(x)g(x)|\,dx \leq \left(\int_a^b |f(x)|^p\,dx\right)^{1/p} \left(\int_a^b |g(x)|^{p'}\,dx\right)^{1/p'}$$

(b). Suppose that $f(x), g(x) \in L[a,b]$ and that there exists some $M > 0$ such that $|g(x)| \leq M$ a.e.. Then $f(x)g(x) \in L[a,b]$ and

$$\int_a^b |f(x)g(x)|\,dx \leq M \int_a^b |f(x)|\,dx.$$

Proof. For (a), take $\{b_k\} \subset\in C[a,b]$ such that $b_k \to f$ in $L^p[a,b]$ and $\{c_k\} \subset\in C[a,b]$ such that $c_k \to g$ in $L^{p'}[a,b]$. We show that $\{b_k c_k\}$ is Cauchy in $L[a,b]$ and that $f(x)g(x)$ is a realization of this sequence.

Note that there exists some $M > 0$ such that for all k, l

$$\|b_l\|_{L^p[a,b]}, \|c_k\|_{L^{p'}[a,b]} \leq M.$$

Then

$$\begin{aligned} &\|b_k c_k - b_l c_l\|_{L[a,b]} \\ \leq &\|\left(b_k - b_l\right) c_k\|_{L[a,b]} + \|b_l \left(c_k - c_l\right)\|_{L[a,b]} \\ \leq &\|\left(b_k - b_l\right)\|_{L^p[a,b]} \|c_k\|_{L^{p'}[a,b]} + \|b_l\|_{L^p[a,b]} \|\left(c_k - c_l\right)\|_{L^{p'}[a,b]} \end{aligned}$$

Since $\|\left(b_k - b_l\right)\|_{L^p[a,b]} \to 0$ and $\|\left(c_k - c_l\right)\|_{L^{p'}[a,b]} \to 0$ as $k, l \to \infty$, we can conclude that $\|b_k c_k - b_l c_l\|_{L[a,b]} \to 0$ as $k, l \to \infty$. Since $b_k(x) \to f(x)$ *a.e.* and $c_k(x) \to g(x)$ *a.e.*, it follows that $b_k(x)c_k(x) \to f(x)g(x)$ *a.e.*, making $f(x)g(x)$ as a realization of the Cauchy sequence $\{b_k c_k\}$.

[2] Since $\int_{[a,b]} \phi$ is defined through a limiting process via an approximate sequence not via a Riemann sum using $\phi(x)$, the sign of $\int_{[a,b]} \phi$ does not automatically follow from $\phi(x) \geq 0$ *a.e.*.

Finally, using

$$\int_a^b |f(x)||g(x)|\,dx = \lim_{k\to\infty}\int_a^b |b_k(x)||c_k(x)|\,dx$$

and

$$\int_a^b |f(x)|^p\,dx = \lim_{k\to\infty}\int_a^b |b_k(x)|^p\,dx$$

as well as

$$\int_a^b |g(x)|^{p'}\,dx = \lim_{k\to\infty}\int_a^b |c_k(x)|^{p'}\,dx$$

it follows from the Hölder's inequality

$$\int_a^b |b_k(x)||c_k(x)|\,dx \le \left(\int_a^b |b_k(x)|^p\,dx\right)^{1/p}\left(\int_a^b |c_k(x)|^{p'}\,dx\right)^{1/p'}$$

that the same inequality holds for f and g.

For (b), the key is how the assumption $|g(x)| \le M$ *a.e.* can be reflected in a Cauchy sequence from $C[a,b]$ approximating g in $L[a,b]$. Let $\{c_k\} \subset C[a,b]$ be a Cauchy sequence for $g \in L[a,b]$. Define

$$c_k^M(x) = \begin{cases} M & \text{if } c_k(x) > M \\ c_k(x) & \text{if } -M \le c_k(x) \le M \\ -M & \text{if } c_k(x) < -M \end{cases}$$

We claim that $\{c_k^M\} \subset C[a,b]$ is a Cauchy sequence equivalent to $\{c_k\}$.

First, observing that

$$|c_k^M(x) - c_l^M(x)| \le |c_k(x) - c_l(x)|$$

holds point wise. It then follows that $\{c_k^M\} \subset C[a,b]$ is a Cauchy sequence in $L[a,b]$. Next, similar to the proof of Proposition 1.2.16, we see that

$$|c_k(x) - c_k^M(x)| = (c_k(x) - M)_+ - (c_k(x) + M)_-$$

and that $(c_k(x) - M)_+ = -(M - c_k(x))_- \to 0$ in $L^p[a,b]$, as $M - g(x) \ge 0$ and $M - c_k(x) \to M - g(x)$ in $L^p[a,b]$; similarly, $(c_k(x) + M)_- \to 0$ in $L^p[a,b]$. This shows that $\{c_k^M\}$ is equivalent to $\{c_k(x)\}$.

Now if $\{b_k\} \subset C[a,b]$ is a Cauchy sequence for f in $L[a,b]$, then we claim that $\{b_k(x)c_k^M(x)\}$ is a Cauchy sequence for $f(x)g(x)$ in $L[a,b]$. This is because

$$\begin{aligned}&\int_a^b |b_k(x)c_k^M(x) - b_l(x)c_l^M(x)|\,dx \\ \le&\int_a^b |\left(b_k(x) - b_l(x)\right)c_k^M(x)|\,dx + \int_a^b |b_l(x)\left(c_k^M(x) - c_l^M(x)\right)|\,dx \\ \le& M\int_a^b |b_k(x) - b_l(x)|\,dx + \int_a^b |b_l(x)\left(c_k^M(x) - c_l^M(x)\right)|\,dx\end{aligned}$$

For any $\epsilon > 0$, first find N such that for all $k, l \ge N$, $M\int_a^b |b_k(x) - b_l(x)|\,dx < \epsilon$. Note that there exists some $M' > 0$ such that $\int_a^b |b_l(x)|\,dx \le M'$ for all l, and that there exists some $\delta > 0$ such that for any open set G with $|G| < \delta$, we have $2M\int_G |b_l(x)|\,dx < \epsilon/M'$ for all l. For such a $\delta > 0$, we can find some open set G with $|G| < \delta$ such that a subsequence of $\{c_k^M(x)\}$, still denoted as $\{c_k^M(x)\}$, satisfies

$c_k^M(x) \to \phi(x)$ uniformly on G^c. Therefore, we can find some K such that for all $k, l \geq K$, we have $|c_k^M(x) - c_l^M(x)| < \epsilon/M'$ for all $x \in G^c$, which leads to

$$\int_{G^c} |b_l(x)\left(c_k^M(x) - c_l^M(x)\right)|\, dx \leq \epsilon.$$

But

$$\int_G |b_l(x)\left(c_k^M(x) - c_l^M(x)\right)|\, dx \leq 2M \int_G |b_l(x)|\, dx < \epsilon.$$

This leads to

$$\int_a^b |b_k(x)c_k^M(x) - b_l(x)c_l^M(x)|\, dx < 3\epsilon.$$

■

Theorem 1.2.18 Monotone Convergence Theorem.

Suppose that $\{\phi_k\}$ is a sequence in $L^1[a,b]$ such that $0 \leq \phi_k(x) \leq \phi_{k+1}(x)$ a.e. for all k and that $\int_a^b \phi_k$ is bounded. Then the a.e. defined $\phi(x) := \lim_{k\to\infty} \phi_k(x)$ (it could take value ∞) is the realization of some $\phi \in L^1[a,b]$, and $\|\phi_k - \phi\|_{L^1[a,b]} \to 0$ as $k \to \infty$.

Proof. First, Proposition 1.2.16 implies that $\int_a^b \phi_k(x)\, dx \leq \int_a^b \phi_{k+1}(x)\, dx$, so the sequence of scalars $\{\int_a^b \phi_k(x)\, dx\}$ as a bounded, monotone increasing sequence is convergent. Since for $l > k$, $\int_a^b |\phi_l(x) - \phi_k(x)|\, dx = \int_a^b (\phi_l(x) - \phi_k(x))\, dx$, it follows that $\{\phi_k\}$ is a Cauchy sequence in $L^1[a,b]$. Then by Proposition 1.2.14 it has a subsequence $\{\phi_{k_l}\}$ converging to some $\tilde{\phi}$ in $L^1[a,b]$, and to $\tilde{\phi}(x)$ *a.e.* pointwise. This identifies $\phi(x) := \lim_{k\to\infty} \phi_k(x)$ with $\tilde{\phi}(x)$ *a.e.* pointwise, showing that $\phi \in L^1[a,b]$. Since the full sequence $\{\phi_k\}$ is a Cauchy sequence in $L^1[a,b]$, we show that $\phi_k \to \phi$ in $L^1[a,b]$ as follows.

For any $\epsilon > 0$, there exists L such that for all $l \geq L$, $\int_a^b |\phi_{k_l}(x) - \phi(x)|\, dx < \epsilon$; also, there exists N such that whenever $k, m \geq N$, $\int_a^b |\phi_k(x) - \phi_m(x)|\, dx < \epsilon$. Let $N' = \max\{N, k_L\}$. Then when $k \geq N'$, we can find l such that $k_l \geq N'$ so

$$\int_a^b |\phi_k(x) - \phi(x)|\, dx \leq \int_a^b |\phi_k(x) - \phi_{k_l}(x)|\, dx + \int_a^b |\phi_{k_l}(x) - \phi(x)|\, dx \leq 2\epsilon,$$

proving that $\phi_k \to \phi$ in $L^1[a,b]$. ■

Corollary 1.2.19

Suppose that G_n is a sequence of open set such that $G_{n+1} \subset G_n$ for all n and $\cap_{n=1}^{\infty} G_n$ is negligible and that $|G_1| < \infty$, then $|G_n| \to 0$ as $n \to \infty$.

Proof. Note that $\chi_{G_{n+1}}(x) \leq \chi_{G_n}(x)$ and that $\chi_{G_n}(x) \to 0$ except on $\cap_{n=1}^{\infty} G_n$. It follows that $1 - \chi_{G_{n+1}}(x) \geq 1 - \chi_{G_n}(x)$ and that $1 - \chi_{G_n}(x) \to 1$ except on $\cap_{n=1}^{\infty} G_n$.

Then Theorem 1.2.18 implies that

$$\int_a^b (1 - \chi_{G_n}(x))\, dx \to \int_a^b 1\, dx = b - a,$$

from which we conclude that $|G_n| = \int_a^b \chi_{G_n}(x)\, dx \to 0$. ■

The assumption that $|G_1| < \infty$ (it suffices that some $|G_n| < \infty$) cannot be removed, for $G_n = (n, \infty)$ is a decreasing sequence of open set with $\cap_{n=1}^{\infty} G_n = \emptyset$ is certainly negligible, but $|G_n|$ does not converge to 0.

Exercise 1.2.20 Suppose that $\{\phi_k\}$ is a sequence in $L^1[a,b]$ such that $\phi_k(x) \geq \phi_{k+1}(x)$ *a.e.* for all k and that $\int_a^b \phi_k$ is bounded. Prove that the *a.e.* defined $\phi(x) := \lim_{k\to\infty} \phi_k(x)$ is the realization of some $\phi \in L^1[a,b]$, and $\|\phi_k - \phi\|_{L^1[a,b]} \to 0$ as $k \to \infty$.

Exercise 1.2.21 Suppose that $\{\phi_k\}$ is a sequence in $L^1[a,b]$ such that $\phi_k(x) \geq 0$ *a.e.* for all k and that $\sum_{k=1}^{l} \int_a^b \phi_k$ is bounded independent of k. Prove that $\lim_{l\to\infty} \sum_{k=1}^{l} \phi_k(x)$ is the realization of some of element in $L^1[a,b]$. Denoting it as $\sum_{k=1}^{\infty} \phi_k$, prove that

$$\int_a^b \left(\sum_{k=1}^{\infty} \phi_k \right) dx = \sum_{k=1}^{\infty} \left(\int_a^b \phi_k \, dx \right).$$

Theorem 1.2.22 Fatou Theorem.

Suppose that $\{\phi_k\}$ is a sequence in $L^1[a,b]$ such that $\phi_k(x) \geq 0$ a.e. for all k and that $\liminf_{k\to\infty} \int_a^b \phi_k(x)$ is finite. Then

$$\liminf_{k\to\infty} \phi_k(x) := \lim_{l\to\infty} \inf\{\phi_k(x) : k \geq l\} \in L^1[a,b]$$

and

$$\int_a^b \liminf_{k\to\infty} \phi_k(x) \leq \liminf_{k\to\infty} \int_a^b \phi_k(x).$$

Proof. Define $\psi_{k,l} = \min\{\phi_k, \cdots, \phi_{k+l}\}$. Then $\psi_{k,l} \in L^1[a,b]$ and $\psi_{k,l}(x) \geq \psi_{k,l+1}(x)$ *a.e.*. Then Theorem 1.2.18 implies that $\Phi_k = \lim_{l\to\infty} \psi_{k,l} \in L^1[a,b]$ with $\int_a^b \Phi_k = \lim_{l\to\infty} \int_a^b \psi_{k,l}$. Since for any $l > l' \geq k$, $\psi_{k,l} \leq \phi_{l'}$, it follows that $\int_a^b \Phi_k \leq \int_a^b \phi_{l'}$, therefore, $\int_a^b \Phi_k \leq \inf\{\int_a^b \phi_{l'} : l' \geq k\}$.

Now $\Phi_k \leq \Phi_{k+1}$ and $\int_a^b \Phi_k$ is bounded under our assumption. Then Theorem 1.2.18 implies $\lim_{k\to\infty} \Phi_k \in L^1[a,b]$ and

$$\int_a^b \liminf_{k\to\infty} \phi_k(x) = \int_a^b \lim_{k\to\infty} \Phi_k = \lim_{k\to\infty} \int_a^b \Phi_k \leq \lim_{k\to\infty} \inf\{\int_a^b \phi_{l'} : l' \geq k\},$$

which equals $\liminf_{k\to\infty} \int_a^b \phi_k(x)$. ■

Exercise 1.2.23 Construct a sequence $\{\phi_k\}$ in $L^1[a,b]$ such that $\phi_k(x) \geq 0$ *a.e.* for all k, and

$$\int_a^b \liminf_{k\to\infty} \phi_k(x) < \liminf_{k\to\infty} \int_a^b \phi_k(x).$$

Theorem 1.2.24 Dominated Convergence Theorem.

Suppose that $\{\phi_k\}$ is a sequence in $L^1[a,b]$ such that $\lim_{k\to\infty} \phi_k(x) = \phi(x)$ exists a.e., and there exists a $g \in L^1[a,b]$ such that $|\phi_k(x)| \leq g(x)$ a.e.. Then $\phi \in L^1[a,b]$ and $\phi_k \to \phi$ in $L^1[a,b]$. As a consequence, $\int_a^b \phi_k \to \int_a^b \phi$ as $k \to \infty$.

Proof. Define $\psi_{k,l} = \min\{\phi_k, \cdots, \phi_{k+l}\}$. Then $\psi_{k,l} \in L^1[a,b]$ and $\psi_{k,l}(x) \geq \psi_{k,l+1}(x)$ *a.e.*. ϕ_k and $\psi_{k,l}$ are no longer necessarily non-negative, but $|\psi_{k,l}(x)| \leq g(x)$ *a.e.*. Then $\{g(x) \pm \psi_{k,l}(x)\}$ are non-negative functions in $L^1[a,b]$, monotone in l. Therefore, $\lim_{l\to\infty}(g(x) \pm \psi_{k,l}(x)) \in L^1[a,b]$, with

$$\int_a^b \left(g \pm \lim_{l\to\infty} \psi_{k,l}\right) = \lim_{l\to\infty} \int_a^b (g \pm \psi_{k,l}),$$

from which it follows that $\Phi_k = \lim_{l\to\infty} \psi_{k,l} \in L^1[a,b]$, with

$$\int_a^b \Phi_k = \lim_{l\to\infty} \int_a^b \psi_{k,l}.$$

Note that it follows from $\phi_k(x) \to \phi(x)$ *a.e.* that $\Phi_k(x) \to \phi(x)$ *a.e.* as $k \to \infty$. Therefore by Theorem 1.2.18 $\phi \in L^1[a,b]$.

We can now apply Fatou Theorem to the sequence of non-negative functions $2g - |\phi_k - \phi| \in L^1[a,b]$ to imply

$$\int_a^b \liminf_{k\to\infty} (2g - |\phi_k - \phi|) \leq \liminf_{k\to\infty} \int_a^b (2g - |\phi_k - \phi|) = \int_a^b 2g - \limsup_{k\to\infty} \int_a^b |\phi_k - \phi|.$$

Since $\liminf_{k\to\infty}(2g - |\phi_k - \phi|) = 2g$, it follows that $\limsup_{k\to\infty} \int_a^b |\phi_k - \phi| = 0$. ■

Second proof in special cases. We provide another proof when each of $\phi_k \in C[a,b]$. Let E be a negligible set such that $\phi_k(x) \to \phi(x)$ on E^c. Recall that

$$E^c = \cap_{m=1}^{\infty} \cup_{N=1}^{\infty} \cap_{k=N}^{\infty} \{x : |\phi_k(x) - \phi(x)| \leq 1/m\}.$$

Therefore,

$$E = \cup_{m=1}^{\infty} \cap_{N=1}^{\infty} \cup_{k=N}^{\infty} \{x : |\phi_k(x) - \phi(x)| > 1/m\}$$

and each $\cap_{N=1}^{\infty} \cup_{k=N}^{\infty} \{x : |\phi_k(x) - \phi(x)| > 1/m\}$ is negligible. Since we have not developed the theory of measurable sets and measure, it's not easy to make use of this information.

Under the assumption that each $\phi_k \in C[a,b]$, we can modify the above relation as

$$E = \cup_{m=1}^{\infty} \cap_{N=1}^{\infty} \cup_{k,l=N}^{\infty} \{x : |\phi_k(x) - \phi_l(x)| > 1/m\}$$

and each $G_{m,N} := \cup_{k,l=N}^{\infty} \{x : |\phi_k(x) - \phi_l(x)| > 1/m\}$ is open. Now that $\cap_{N=1}^{\infty} G_{m,N}$ is negligible, it follows from Corollary 1.2.19 that $|G_{m,N}| \to 0$ as $N \to \infty$. Using $g \in L[a,b]$, it follows that, for any $\epsilon > 0$, there exists some $\delta > 0$ that for any open set G in $[a,b]$ with $|G| < \delta$, we have $\int_G g\, dx < \epsilon$. It then follows that $\int_G |\phi_k(x)|\, dx < \epsilon$ and $\int_G |\phi(x)|\, dx \leq \epsilon$, therefore $\int_G |\phi_k(x) - \phi(x)|\, dx \leq 2\epsilon$. It further follows that, there exists N_m such that $|G_{m,N_m}| < \delta/2^m$, from which we deduce that the open set $G := \cup_{m=1}^{\infty} G_{m,N_m}$ satisfies $|G| < \delta$ and in G^c we have $|\phi_k(x) - \phi_l(x)| \leq 1/m$ for $k, l \geq N_m$. This shows that $\{\phi_k\}$ converges (to ϕ) uniformly on G^c.

It now follows that, for the G above, $\int_G |\phi_k(x) - \phi(x)|\, dx \leq 2\epsilon$, and using $\{\phi_k\}$ converges (to ϕ) uniformly on G^c, we argue that $\int_{G^c} |\phi_k(x) - \phi(x)|\, dx \to 0$. This shows that $\int_a^b |\phi_k(x) - \phi(x)|\, dx \to 0$. ■

Exercise 1.2.25 Suppose that $\{\phi_k\}$ is a sequence in $L^1[a,b]$ and that $\sum_{k=1}^{\infty} \int_a^b |\phi_k|$ is convergent. Prove that $\lim_{l\to\infty} \sum_{k=1}^{l} \phi_k(x)$ is the realization of some of element in $L^1[a,b]$. Denoting it as $\sum_{k=1}^{\infty}\phi_k$, prove that

$$\int_a^b\left(\sum_{k=1}^{\infty}\phi_k\right)\,dx=\sum_{k=1}^{\infty}\left(\int_a^b\phi_k\,dx\right).$$

1.3 The Space of Riemann Integrable Functions

Here we discuss the relation between the space of Riemann integrable functions on $[a,b]$ and $L^1[a,b]$, the completion of $C[a,b]$ under the $L^1[a,b]$ norm.

We first make the following definition.

Definition 1.3.1

Let f be defined on an interval I. The oscillation of a function f over the set $S\subset I$ is defined to be

$$\sup_S f-\inf_S f=M(f,S)-m(f,S)$$

and is denoted as $\operatorname{osc}(f,S)$.

The oscillation of a function f at a point x is defined to be

$$\lim_{r\searrow 0}\operatorname{osc}(f,I\cap I(x,r)),$$

and is denoted as $\operatorname{osc}(f)(x)$. Here $I(x,r)$ is the open interval of radius r centered at x.

Proposition 1.3.2 Upper Semi-continuity of $\operatorname{osc}(f)(x)$.

The function $osc(f)(x)$ is upper semi-continuous. As a consequence, for any real number a, the set $\{x:osc(f)(x)\ge a\}$ is closed.

Proof. For any real number a, if $x_0\in\{x:\operatorname{osc}(f)(x)<a\}$, then there exists some $r>0$ such that $\operatorname{osc}(f,I(x_0,r))<a$. For any $x\in I(x_0,r)$, we observe that $\operatorname{osc}(f)(x)\le\operatorname{osc}(f,I(x_0,r))<a$. Thus $I(x_0,r)\subset\{x:\operatorname{osc}(f)(x)<a\}$, proving that the latter is open. ■

Note that

$$\{x:f\text{ discontinuous at }x\}=\cup_{k\in\mathbb{N}}\{y:\operatorname{osc}(f)(y)\ge\frac{1}{k}\},\tag{1.3.1}$$

namely, $\{x:f$ discontinuous at $x\}$ is a countable union of of the closed sets $\{y:\operatorname{osc}(f)(y)\ge\frac{1}{k}\}$.

Theorem 1.3.3 Riemann Integrability Criterion in terms of the Oscillation of the Integrand.

A bounded real-valued function f defined on the interval I is Riemann integrable iff

$$\forall\epsilon>0,\exists\text{ a partition }\mathcal{P}:=\{I_\alpha\}\text{ such that }\sum_\alpha osc(f,I_\alpha)|I_\alpha|<\epsilon.\tag{1.3.2}$$

In particular, if f is continuous on the closed interval I, then f is

Riemann integrable on I.

Proof. Suppose that f is Riemann integrable on I and that $\epsilon > 0$ is given. Then the Darboux integrability criterion gives us a partition $\mathcal{P} := \{I_\alpha\}$ such that

$$\int_I f - \frac{\epsilon}{2} < L(f, \mathcal{P}) \leq \int_I f \leq U(f, \mathcal{P}) < \int_I f + \frac{\epsilon}{2},$$

which implies that

$$\sum_\alpha \operatorname{osc}(f, I_\alpha)|I_\alpha| = U(f, \mathcal{P}) - L(f, \mathcal{P}) < \epsilon.$$

Suppose that (1.3.2) holds. Then $U(f, \mathcal{P}) - L(f, \mathcal{P}) < \epsilon$, and

$$0 \leq \overline{\int_I} f - \underline{\int_I} f \leq U(f, \mathcal{P}) - L(f, \mathcal{P}) < \epsilon.$$

Since $\epsilon > 0$ is arbitrary, it follows that $\overline{\int_I} f - \underline{\int_I} f = 0$, and f is Riemann integrable on I.

Finally, suppose that f is continuous on the closed interval I, then it is uniformly continuous on I. For any given $\epsilon > 0$, there exists $\delta > 0$ such that for any partition $\mathcal{P} := \{I_\alpha\}$ of I with $\lambda(P) < \delta$, we have $\operatorname{osc}(f, I_\alpha) < \epsilon/|I|$ for all $I_\alpha \in \mathcal{P}$. It then follows that

$$\sum_\alpha \operatorname{osc}(f, I_\alpha)|I_\alpha| \leq \epsilon,$$

proving the Riemann integrability of f on I. ■

Definition 1.3.4

A set $S \subset \mathbb{R}$ is said to have content 0, if for any $\epsilon > 0$, there exists a *finite* cover $\{I_i\}$ of S by intervals such that

$$\sum_i |I_i| < \epsilon.$$

We are now ready to formulate the following theorem.

Theorem 1.3.5 Riemann Integrability in terms of the Set of Discontinuity.

A bounded function f on a bounded closed interval I is Riemann integrable on I iff its set of discontinuity is negligible.

Proof. We will use (1.3.1) for the only if part.

Suppose that f is Riemann integrable on I. For each $k \in \mathbb{N}$, we will prove that the closed set $D_k := \{y : \operatorname{osc}(f)(y) \geq \frac{1}{k}\}$ is a set of content 0.

Given any $\epsilon > 0$. There exists a partition $\mathcal{P} = \{I_\alpha\}$ of I such that

$$\sum_\alpha \operatorname{osc}(f, I_\alpha)|I_\alpha| < \frac{\epsilon}{2k}.$$

The intervals in $\mathcal{P}$ are divided into two subgroups: the subgroup $\mathcal{L}_k$ consisting those I_α such that $\operatorname{osc}(f, I_\alpha) \geq \frac{1}{2k}$, and the subgroup $\mathcal{S}_k$ consisting those I_α such that

$\operatorname{osc}(f, I_\alpha) < \frac{1}{2k}$. Then it follows from

$$\frac{1}{2k} \sum_{I_\alpha \in \mathcal{L}_k} |I_\alpha| \leq \sum_{I_\alpha \in \mathcal{L}_k} \operatorname{osc}(f, I_\alpha)|I_\alpha| < \frac{\epsilon}{2k}$$

that

$$\sum_{I_\alpha \in \mathcal{L}_k} |I_\alpha| < \epsilon.$$

We now claim that

$$D_k \subset \cup_{I_\alpha \in \mathcal{L}_k} I_\alpha.$$

This will show that D_k is a set of content 0.

If the claim were not true, there would exist some $x \in D_k \setminus \cup_{I_\alpha \in \mathcal{L}_k} I_\alpha$. Thus $x \in \cup_{I_\alpha \in \mathcal{S}_k} I_\alpha$. Since the complement of $\cup_{I_\alpha \in \mathcal{L}_k} I_\alpha$ is open, there exists some interval $I(x,r) \subset \cup_{I_\alpha \in \mathcal{S}_k} I_\alpha$. If $x \in$ interior(I_α) for some $I_\alpha \in \mathcal{S}_k$, it would force $\frac{1}{k} \leq \operatorname{osc}(f)(x) \leq \operatorname{osc}(f, I_\alpha) < \frac{1}{2k}$, which would be a contradiction. So x can only be on the boundary of one or more $I_\alpha \in \mathcal{S}_k$. We can choose $r > 0$ small enough such that any $y \in I(x,r)$ and x will be in one such common interval. Therefore, $|f(y) - f(x)| < \frac{1}{2k}$. This would lead to $\operatorname{osc}(f)(x) \leq \operatorname{osc}(f, I(x,r)) < \frac{1}{k}$, contradicting $x \in D_k$.

For the if part, take any $\epsilon > 0$, then choose k such that $k^{-1} < \epsilon$. The closed set $D_k = \{y : \operatorname{osc}(f)(y) \geq \frac{1}{k}\}$ is a subset of the negligible set D, so can be covered by a union G of a finite number of intervals such that $|G| < \epsilon$. Every point $x \in [a,b] \setminus G$ has an open interval $I(x,r)$ such that $\operatorname{osc}(f, I(x,r)) < k^{-1}$. The closed set $[a,b] \setminus G$ can be covered by the union of a finite number of such intervals and one can find a fine enough partition $\mathcal{P} = \{I_i\}$ of $[a,b]$ such that if any I_i contains a point of D_k^c, then $\operatorname{osc}(f, I_i) < k^{-1}$; and the union of those $I_i \subset D_k$ has their sum of lengths $< \epsilon$. This would give

$$\sum_i \operatorname{osc}(f, I_i)|I_i| \leq k^{-1}(b-a) + 2\epsilon \sup_{[a,b]} |f|,$$

which proves that f satisfies the Riemann integrability criterion. ■

Corollary 1.3.6 Approximation Property of Functions in $\mathcal{R}[a,b]$.

Any Riemann integrable function $f \in \mathcal{R}[a,b]$ lies in $L^p[a,b]$ for any $1 \leq p < \infty$, and for any $\epsilon > 0$ there exists a continuous function $c \in C[a,b]$ and an open set G in $[a,b]$ such that $|G| < \epsilon$, $f = c$ on G^c, and $\max_{[a,b]} |c| \leq \max_{[a,b]} |f|$.

Proof. Since the set D of discontinuity of f is negligible, for any $\epsilon > 0$ there exists an open set G in $[a,b]$ such that it covers D and $|G| < \epsilon$. For every point x in the closed set $[a,b] \setminus G$, $\operatorname{osc}(f)(x) = 0$, so f is a continuous function on this closed set. It can be extended to a continuous function c on $[a,b]$ such that $\max_{[a,b]} |c| \leq \max_{[a,b]} |f|$. Then this c satisfies our requirements. ■

Exercise 1.3.7 Suppose that $c(x) \in C[a,b]$ and $E \subset [a,b]$ is negligible. Is it true that any modification of $c(x)$ on E into a bounded function $\tilde{c}(x)$ will remain to have a negligible set of discontinuity and remain Riemann integrable on $[a,b]$?

Exercise 1.3.8 Suppose that E is a closed negligible subset of $[a,b]$. Prove that the set of discontinuity of the characteristic function $\chi_E(x)$ is precisely E.

Exercise 1.3.9 Suppose that E is the union of at most countably many closed negligible subsets of $[a,b]$. Prove that there is a bounded function $f(x)$ defined on $[a,b]$ whose set of discontinuity is precisely E.

1.4 Extension of the Fundamental Theorem of Calculus; Absolutely Continuous Functions

Corollary 1.1.12 extends the Fundamental Theorem of Calculus to those monotone functions α such that α' exists and is Riemann integrable. Here we discuss its further generalization and related issues. Questions that need to be addressed include

(I). Identify conditions that guarantee that α' exists and is Riemann integrable, or *a.e.* exists and is in $L^1[a,b]$.

(II). Does (1.1.10) hold for any function α such that α' *a.e.* exists and is in $L^1[a,b]$? More generally, under what conditions on a function f on $[a,b]$ there would exist a function $g \in L^1[a,b]$ such that

$$f(t) - f(a) = \int_a^t g(x)\,dx \tag{1.4.1}$$

holds?

The last question has two parts: (a). The necessary conditions for (1.4.1) to hold, in particular, does it hold that $f'(x) = g(x)$ for $a.e. x \in [a,b]$? (b). The sufficient conditions for (1.4.1) to hold.

Definition 1.4.1 Absolute Continuous Function.

A function f defined on $[a,b]$ is said to be **absolutely continuous** on $[a,b]$ if for any $\epsilon > 0$, there exists a $\delta > 0$ such that for any finite collection of disjoint intervals $\{[a_i,b_i] : 1 \le i \le k\}$ of $[a,b]$ with its total length $\sum_{i=1}^k (b_i - a_i) < \delta$,

$$\sum_{i=1}^k |f(b_i) - f(a_i)| < \epsilon. \tag{1.4.2}$$

By Exercise 1.2.3, if (1.4.1) holds, then f is absolutely continuous on $[a,b]$.

If f is **Lipschitz continuous** on $[a,b]$, namely, there exists some $L > 0$ such that $|f(x) - f(y)| \le L|y - x|$ for any $x, y \in [a,b]$, then f is absolutely continuous on $[a,b]$.

If $f \in C[a,b]$ is differentiable everywhere in (a,b) and f' is bounded on (a,b), then f is Lipschitz continuous on $[a,b]$, therefore, is absolutely continuous on $[a,b]$.

Cantor's function is differentiable *a.e.* in (a,b), yet is not absolutely continuous on $[a,b]$.

Exercise 1.4.2 Suppose that $f \in C[a,b]$ is differentiable *a.e.* in (a,b) and $|f'| \le M < \infty$ *a.e.* on (a,b). Is f necessarily Lipschitz continuous on $[a,b]$? Is it necessarily absolutely continuous on $[a,b]$?

We will prove below that any absolutely continuous function on $[a,b]$ is differentiable *a.e.* on $[a,b]$ with its derivative a function in $L^1[a,b]$. Assuming this for now, we prove

Theorem 1.4.3 Differentiability of an Indefinite Integral.

Suppost that $g \in L^1[a,b]$. Then the function $f(x) := \int_a^x g(t)\,dt$ is differentiable a.e. on $[a,b]$ and $f'(x) = g(x) a.e.$ on $[a,b]$.

Proof. According to our discussion above, $f(x) := \int_a^x g(t)\,dt$ is differentiable *a.e.* on $[a,b]$ and $f'(x) \in L^1[a,b]$. This implies that $g_h(x) := h^{-1}\int_x^{x+h} g(t)\,dt = h^{-1}(f(x+h) - f(x)) \to f'(x)$ *a.e.* on $[a,b]$ as $|h| \to 0$.

On the other hand, according to Exercise 1.2.12, $\int_a^b |g_h(x) - g(x)|\,dx \to 0$ as $h \to 0$. Then by (i) of Proposition 1.2.8, there exists a sequence $h_k \to 0$ such that $g_{h_k}(x) \to g(x)$ *a.e.* on $[a,b]$ as $k \to \infty$. This then identifies $f'(x) = g(x)$ *a.e.* on $[a,b]$.

We could also complete the last part by using the Fatou Theorem:

$$\begin{aligned}\int_a^b |f'(x) - g(x)|\,dx &= \int_a^b \lim_{k\to\infty} |g_{h_k}(x) - g(x)|\,dx \\ &\le \liminf k \to \infty \int_a^b |g_{h_k}(x) - g(x)|\,dx = 0,\end{aligned}$$

from which we conclude that $f'(x) = g(x)$ *a.e.* on $[a,b]$ via (ii) of Proposition 1.2.8. ■

For the remaining part, we will state and use Lebesgue's differentiability theorem of a monotone function but will skip its proof.

Theorem 1.4.4 Lebesgue's Differentiability Theorem of a Monotone Function.

Any monotone function on $[a,b]$ is differentiable a.e. on $[a,b]$.

Proposition 1.4.5 Integrability of the Derivative of a Monotone Function.

Suppose that f is a monotone increasing function on $[a,b]$, then its derivative, $f'(x)$, is the realization of a function in $L^1[a,b]$. Furthermore,

$$\int_a^b f'(x)\,dx \le f(b) - f(a).$$

Proof. According to Theorem 1.4.4, $h^{-1}(f(x+h) - f(x)) \to f'(x)a.e.$ on $[a,b]$. For each $h > 0$, $h^{-1}(f(x+h) - f(x))$ is a non-negative function in $\mathcal{R}[a,b]$ (extend $f(x) = f(b)$ for $x > b$), and

$$\int_a^b h^{-1}(f(x+h) - f(x))\,dx = h^{-1}\left(\int_b^{b+h} f(x)\,dx - \int_a^{a+h} f(x)\,dx\right) \le f(b) - f(a),$$

using $\int_a^{a+h} f(x)\,dx \ge f(a)h$. Thus by Fatou Theorem,

$$\begin{aligned}\int_a^b f'(x)\,dx &= \int_a^b \liminf_{h\to 0+} h^{-1}(f(x+h) - f(x))\,dx \\ &\le \liminf_{h\to 0+} \int_a^b h^{-1}(f(x+h) - f(x))\,dx \le f(b) - f(a).\end{aligned}$$

■

Definition 1.4.6 Function of Bounded Variation.

A function f on $[a,b]$ is said to have bounded variation on $[a,b]$, if there exists some $M > 0$ such that for any partition $\mathcal{P} = \{a = a_0 < a_1 < \cdots < a_k = b\}$,

$$\sum_{i=1}^{k} |f(a_i) - f(a_{i-1})| \le M.$$

$\sup_{\mathcal{P}} \sum_{i=1}^{k} |f(a_i) - f(a_{i-1})|$ is called the total variation of f on $[a, b]$ and is denoted as $V(f, [a, b])$.

If f is a monotone increasing function on $[a, b]$, then it has bounded variation on $[a, b]$ and $V(f, [a, b]) = f(b) - f(a)$.

If $f = g - h$ is the difference of two monotone increasing functions on $[a, b]$, then it has bounded variation on $[a, b]$ and $V(f, [a, b]) \le V(g, [a, b]) + V(h, [a, b])$.

Proposition 1.4.7

Any function of bounded variation on $[a, b]$ is the difference of two monotone increasing functions.

Proof. Let f be a function of bounded variation on $[a, b]$. Define $g(x) := V(f, [a, x])$. Then $g(x)$ is a monotone increasing function on $[a, b]$. We now prove that $h(x) := g(x) - f(x)$ is a monotone increasing function on $[a, b]$.

Take $x < y$ in $[a, b]$, then

$$h(y) - h(x) = g(y) - g(x) - [f(y) - f(x)] = V(f, [x, y]) - [f(y) - f(x)].$$

$V(f, [x, y])$ is the supremum of $\sum_{i=1}^{k} |f(a_i) - f(a_{i-1})|$ for any partition $\{x = a_0 < a_1 < \cdots < a_k = y\}$, so $V(f, [x, y]) \ge |f(y) - f(x)|$. ■

An absolutely continuous function on $[a, b]$ has bounded variation on $[a, b]$, therefore, according to Proposition 1.4.7, Theorem 1.4.4 and Proposition 1.4.5, has its derivative in $L^1[a, b]$.

Theorem 1.4.8 Fundamental Theorem of Calculus.

A function f on $[a, b]$ satisfies (1.4.1) for some function $g \in L^1[a, b]$ iff f is absolutely continuous on $[a, b]$. When (1.4.1) holds, $f'(x) = g(x)a.e.$ on $[a, b]$.

Proof. We already discussed the necessary part. For the sufficient part, suppose that f is absolutely continuous on $[a, b]$. Then, according to our discussion above, $f'(x)$ exists *a.e.* on $[a, b]$ and is an element in $L^1[a, b]$. Set $F(x) = \int_a^x f'(t)\, dt$. Then, according to Theorem 1.4.3, $F(x)$ is is absolutely continuous on $[a, b]$ and $F'(x) = f'(x)$ *a.e.* on $[a, b]$.

Now $f(x) - F(x)$ is absolutely continuous on $[a, b]$ and $(f(x) - F(x))' = 0$ *a.e.* on $[a, b]$. We can draw our conclusion based on the following Lemma. ■

Lemma 1.4.9 Absolute Function with Vanishing Derivative.

Suppose that $g(x)$ is absolutely continuous on $[a, b]$ and $g'(x) = 0a.e.$ on $[a, b]$. Then g must be a constant function on $[a, b]$.

We will use the concept and properties of Vitali covering in proving this.

Definition 1.4.10 Vitali Covering.

Suppose that $\Gamma = \{I_\alpha\}$ is a family of intervals covering a set $E \subset (a, b)$. Suppose that for any $\epsilon > 0$ and any $x \in E$, there exists an interval $I \in \Gamma$ such that $x \in I$ and $|I| < \epsilon$. Then Γ is said for form a Vitali covering of E

Lemma 1.4.11 Vitali Covering Lemma.

Suppose that $\Gamma = \{I_\alpha\}$ is a Vitali covering of $E \subset (a, b)$. Then for any $\epsilon > 0$, there exists a finite collection of disjoint intervals $\{I_i, 1 \le i \le k\}$ of Γ such that $E \setminus \cup_{i=1}^k I_i$ can be covered by an open set G with $|G| < \epsilon$.

Proof. We may take the I_α to be closed intervals in (a, b). We choose I_i inductively. Choose I_1 from Γ such that $2|I_1| > \sup |I_\alpha|$. After $\{I_i, 1 \le i \le l\}$ are chosen, if $E \setminus \cup_{i=1}^l I_i \ne \emptyset$, define $\delta_{l+1} = \sup\{|I_\alpha| : I_\alpha \cap I_i = \emptyset, i = 1, \cdots, l\}$. Then $\delta_{l+1} > 0$ and we choose some $I_{l+1} \in \Gamma$ such that $I_{l+1} \cap I_i = \emptyset, i = 1, \cdots, l$ and $2|I_{l+1}| > \delta_{l+1}$.

Since this collection of disjoint intervals are all contained in (a, b), $\sum_l |I_l| < \infty$. Thus, for a given $\epsilon > 0$, there exists k such that $\sum_{l>k} |I_l| < \epsilon/5$.

Any $x \in E \setminus \cup_{i=1}^l I_i$ must be contained in some $I \in \Gamma$ such that $I \cap I_i = \emptyset, i = 1, \cdots, k$. Since $|I| > 0$ and $\delta_l \to 0$ as $l \to \infty$, we claim that $I \cap I_l \ne \emptyset$ for some $l > k$, for, whenever $\delta_{l'} < |I|$, $I \cap I_l \ne \emptyset$ for some $l < l'$. Let I'_l be the interval with the same center point as I_l but $|I'_l| = 5|I_l|$, then $I \subset I'_l$ whenever $I \cap I_l \ne \emptyset$. Thus $\cup_{l>k} I'_l$ covers $E \setminus \cup_{i=1}^l I_i$ with $\sum_{l>k} |I'_l| = 5 \sum_{l>k} |I_l| < \epsilon$. ■

Proof of Lemma 1.4.9. Suppose g is not a constant function on $[a, b]$. Then there exists some $c \in (a, b)$ such that $|g(c) - g(a)| > 2\epsilon$ for some $\epsilon > 0$. Consider $E_c := \{x \in (a, c) : g'(x) = 0\}$. Then $[a, c] \setminus E_c$ is negligible. For any $x \in E_c$ and any $r > 0$, for all sufficiently small $h > 0$, we have $[x, x + h] \subset (a, c)$ and

$$|g(x + h) - g(x)| \le rh.$$

Thus for any fixed such $r > 0$, the family $\{[x, x + h] : x \in E_c\}$ forms a Vitali cover of E_c. Thus, for any $\delta > 0$, there exists disjoint intervals

$$[x_1, x_1 + h_1], \cdots, [x_k, x_k + h_k]$$

such that

$$[a, c] \setminus \cup_{i=1}^k [x_i, x_i + h_i] \subset ([a, c] \setminus E_c) \cup \left(E_c \setminus \cup_{i=1}^k [x_i, x_i + h_i]\right)$$

can be covered by an open set G with $|G| < \delta$.

We may assume that

$$a < x_1 < x_1 + h_1 < x_2 < x_2 + h_2 < \cdots < x_k < x_k + h_k < c.$$

These points, together with a and c, forms a partition of $[a, c]$. The sum of the lengths of those intervals of this partition not-overlapping with any of $(x_i, x_i + h_i)$ is $< \delta$. Then

$$2\epsilon < |g(c) - g(a)| \le \sum_{i=1}^k |g(x_i) - g(x_i + h_i)| + \sum_{i=0}^k |g(x_i + h_i) - g(x_{i+1})|$$

where we set $x_0 + h_0 = a$ and $x_{k+1} = c$.

We fix $r > 0$ such that $r(b - a) < \epsilon$. Using the absolute continuity of f on $[a, b]$, there exists some $\delta > 0$ such that for any collection of disjoint intervals

$[a_j, b_j]$ with $\sum_j (b_j - a_j) < \delta$, we have $\sum_j |g(b_j) - g(a_j)| < \epsilon$. For this reason, $\sum_{i=0}^{k} |g(x_i + h_i) - g(x_{i+1})| < \epsilon$. But

$$\sum_{i=1}^{k} |g(x_i) - g(x_i + h_i)| \leq \sum_{i=1}^{k} rh_i \leq r(b - a) < \epsilon.$$

We now see a contradiction and must conclude that g is a constant function on $[a, b]$ ■

Exercise 1.4.12 Suppose that f is absolutely continuous on (a, b) and $|f'| \leq M < \infty$ *a.e.* on (a, b). Is f necessarily Lipschitz continuous on $[a, b]$?

Exercise 1.4.13 Suppose that f has bounded variation on $[0, 1]$ and is absolutely continuous on $[\epsilon, 1]$ for any $0 < \epsilon < 1$. Furthermore, suppose that $f(x)$ is continuous at $x = 0$. Prove that f is absolutely continuous on $[0, 1]$.

Chapter 2

Sequences and Series of Functions

Taking the limit of a sequence of functions is a most natural and necessary process in analysis, and it is an important way of how new classes of functions arise. It is crucial to understand whether and when certain properties of the functions in the sequence (such as continuity or integrability) will pass to the limit function. The question is also equivalent to whether two different limit processes applied to a sequence of functions can be *exchanged*.

The main questions that concern us in this chapter are the following. Suppose that $\{f_n(x)\}$ is a sequence of functions defined on a set E, and for any $x \in E$, $f_n(x) \to f(x)$ as $n \to \infty$. Some of the questions can be raised for fairly general E, but one may first restrict to the case when E is an interval of $\mathbb{R}$.

(a). If each $f_n(x)$ is continuous on E, does it imply that $f(x)$ is continuous on E? If not, what kind of "bad" behavior can $f(x)$ exhibit? What additional conditions on the convergence would guarantee that $f(x)$ is continuous on E? Here we formulate the question using continuity over E, but we could also formulate the question using continuity at a specific point $\mathbf{x}_0$ in E. Then answer to the question depends on whether we can exchange the following two limits to get an equality:

$$\lim_{\mathbf{x}\to\mathbf{x}_0} \lim_{n\to\infty} f_n(\mathbf{x}) = \lim_{n\to\infty} \lim_{\mathbf{x}\to\mathbf{x}_0} f_n(\mathbf{x}).$$

(b). If each $f_n(x)$ is integrable on E, does it imply that $f(x)$ is also integrable on E? If so, does it imply

$$\lim_{n\to\infty} \int_E f_n(x)\,dx = \int_E \lim_{n\to\infty} f_n(x)\,dx? \tag{2.0.1}$$

If not, what additional conditions on the convergence would guarantee that (2.0.1) holds?

2.1 Concept of Uniform Convergence

Simple examples illustrate that *pointwise limit* of a sequence of continuous functions can fail to be continuous, and that even pointwise limit of a sequence of continuously differentiable functions can also fail to be continuous. The key cause is that the convergence may not be *uniform*, namely, for each $\epsilon > 0$ and each $x \in E$, there exists N which may depend on ϵ as well as on x such that when $n \geq N$, we have $|f_n(x) - f(x)| < \epsilon$; but the minimum N needed for $x \in E$ may not be uniform over $x \in E$. The following definition defines the notion of uniform convergence.

Definition 2.1.1 Uniform Convergence.

Let $\{f_n(x)\}$ be a sequence of functions defined on a set E. Then $\{f_n(x)\}$ converges uniformly on E to a limit function $f(x)$ on E if, for every $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that $|f_n(x) - f(x)| < \epsilon$ for all $n \geq N$ and for *all* $x \in E$.

Remark 2.1.2

Note that one needs to specify the set E when discussing the notion of uniform convergence. For example, the sequence of functions $f_n(x) = x^n$ is converging pointwise to $f(x) := 0$ on $[0,1)$, but not uniformly on $[0,1)$. However, for any fixed $0 < \delta < 1$, $f_n(x) \to 0$ *uniformly over* $[0, \delta]$.

Example 2.1.3 Examine the notion of pointwise and uniform convergence.

1. We continue to work with $f_n(x) = x^n$ and consider it on $[0,1]$. It still converges pointwise, but the limiting function

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

fails to be continuous at $x = 1$. This is due to the failure of uniform convergence on $[0,1]$.

Here we still have

$$\lim_{n\to\infty} \int_0^1 x^n \, dx = \int_0^1 f(x) \, dx.$$

2. If we modify $f_n(x)$ into

$$g_n(x) = (n+1)x^n(1-x).$$

Then $g_n(x) \to g(x) := 0$ pointwise on $[0,1]$, although the convergence is still not uniform, as $\max_{[0,1]} g_n(x) = \left(\frac{n}{n+1}\right)^n \to e^{-1}$ as $n \to \infty$. When we examine $\int_0^1 g_n(x)\, dx$, we find it equal to $(n+2)^{-1} \to 0$ as $n \to \infty$.

3. However, if we modify $g_n(x)$ into $h_n(x) := (n+2)g_n(x)$, we continue to have $h_n(x) \to 0$ pointwise on $[0,1]$. The convergence is still not uniform, as $\max_{[0,1]} h_n(x) = (n+2)\left(\frac{n}{n+1}\right)^n \to \infty$. And we find that

$$\int_0^1 h_n(x) = 1 \not\to \int_0^1 0 \, dx!$$

Note that in all the cases above, once we fix some $0 < \delta < 1$, the relevant sequences of functions converge to 0 uniformly over $[0, \delta]$, so they fail to converge uniformly only over a neighborhood of $x = 1$. To examine question (b) raised earlier, we just need to answer whether the following holds:

$$\forall\, \epsilon > 0, \exists\, 0 < \delta < 1 \text{ and } N \text{ such that for all } n \geq N,$$
$$\left| \int_\delta^1 k_n(x)\, dx \right| < \epsilon,$$

where $(k_n(x))$ stands for one of the sequences $\{f_n(x)\}, \{g_n(x)\}, \{h_n(x)\}$ above. Through direct examination we find that this holds for $\{f_n(x)\}, \{g_n(x)\}$, but fails for $\{h_n(x)\}$.

Remark 2.1.4

In most contexts we confine to $\mathbb{R}$ or $\mathbb{C}$-valued functions defined on a set E. The notion of uniform convergence can be defined as along as there is a quantitative way to describe how close $f_n(x)$ is to $f(x)$. If Y is a metric space with d as the metric, and $f_n, f : E \mapsto Y$, then the notion of uniform convergence of f_n makes sense. More specifically, $f_n \to f$ uniformly over E, if for any $\epsilon > 0$, there exists N such that for all $n \geq N$, $\sup_{x\in E} d(f_n(x), f(x)) < \epsilon$. This notion does not require E to be a metric space.

On the other hand, the notion of uniform continuity *also requires a mechanism of quantitatively describing how close x, y are in E, and it would make sense if both E and Y are metric spaces.*

2.2 Properties of A Uniformly Convergent Sequence of Functions

Theorem 2.2.1 The Limit of A Uniformly Convergent Sequence of Continuous Functions is Continuous.

Suppose that $\{f_n(\mathbf{x})\}$ is a sequence of functions defined on a set E which converges uniformly on E to a limit function $f(\mathbf{x})$ on E, and that $\mathbf{x}_0 \in E$ is such that $\lim_{\mathbf{x}\in E, \mathbf{x}\to\mathbf{x}_0} f_n(\mathbf{x}) := L_n$ exists for each n. Then both $\lim_{\mathbf{x}\in E, \mathbf{x}\to\mathbf{x}_0} f(\mathbf{x})$ and $\lim_{n\to\infty} L_n$ exist and equal each other. Namely

$$\lim_{\mathbf{x}\in E, \mathbf{x}\to\mathbf{x}_0} \lim_{n\to\infty} f_n(\mathbf{x}) = \lim_{n\to\infty} \lim_{\mathbf{x}\in E, \mathbf{x}\to\mathbf{x}_0} f_n(\mathbf{x}).$$

As a consequence, suppose that $\{f_n(\mathbf{x})\}$ is a sequence of continuous functions defined on a set E which converges uniformly on E to a limit function $f(\mathbf{x})$ on E, then $f(\mathbf{x})$ is continuous on E.

Proof. We will first show that $\{L_n\}$ is Cauchy as $n \to \infty$ and that $\{f(\mathbf{x})\}$ is Cauchy as $\mathbf{x} \to \mathbf{x}_0$.

For any $\epsilon > 0$, by the uniform convergence assumption, there exists some $N \in \mathbb{N}$ such that

$$|f_n(\mathbf{x}) - f(\mathbf{x})| < \epsilon \text{ and } |f_n(\mathbf{x}) - f_N(\mathbf{x})| < \epsilon \text{ for all } x \in E, n \geq N.$$

Using $\lim_{\mathbf{x}\in E, \mathbf{x}\to\mathbf{x}_0} f_N(\mathbf{x}) := L_N$, there exists some neighborhood $B(\mathbf{x}_0)$ of $\mathbf{x}_0$ such that

$$|f_N(\mathbf{x}) - L_N| < \epsilon \text{ for all } x \in B(\mathbf{x}_0).$$

Then for all $x \in B(\mathbf{x}_0), n \geq N$, we have

$$|f_n(\mathbf{x}) - L_N| \leq |f_n(\mathbf{x}) - f_N(\mathbf{x})| + |f_N(\mathbf{x}) - L_N| < 2\epsilon.$$

Using $\lim_{\mathbf{x}\in E, \mathbf{x}\to\mathbf{x}_0} f_n(\mathbf{x}) := L_n$ we obtain

$$|L_n - L_N| \leq 2\epsilon \text{ for all } n \geq N.$$

This shows that $\{L_n\}$ is Cauchy as $n \to \infty$, so has a limit. Call it L.

In the same setting, for $\mathbf{x}, \mathbf{y} \in B(\mathbf{x}_0)$,

$$|f(\mathbf{x})-f(\mathbf{y})| \leq |f(\mathbf{x})-f_N(\mathbf{x})|+|f_N(\mathbf{x})-f_N(\mathbf{y})|+|f_N(\mathbf{y})-f(\mathbf{y})| \leq 2\epsilon+|f_N(\mathbf{x})-f_N(\mathbf{y})|.$$

But $|f_N(\mathbf{x}) - f_N(\mathbf{y})| \leq |f_N(\mathbf{x}) - L_N| + |L_N - f_N(\mathbf{y})| \leq 2\epsilon$, so we have $|f(\mathbf{x}) - f(\mathbf{y})| \leq 4\epsilon$, for $\mathbf{x}, \mathbf{y} \in B(\mathbf{x}_0)$, which shows that $\{f(\mathbf{x})\}$ is Cauchy as $\mathbf{x} \to \mathbf{x}_0$ and $\lim_{\mathbf{x}\in E, \mathbf{x}\to\mathbf{x}_0} f(\mathbf{x})$ exists.

Now for any $n \geq N$, sending $\mathbf{x} \to \mathbf{x}_0$ in $|f_n(\mathbf{x}) - f(\mathbf{x})| < \epsilon$, we get

$$|L_n - \lim_{\mathbf{x}\in E, \mathbf{x}\to\mathbf{x}_0} f(\mathbf{x})| \leq \epsilon.$$

Then sending $n \to \infty$ in the above we get

$$|\lim_{n\to\infty} L_n - \lim_{\mathbf{x}\in E, \mathbf{x}\to\mathbf{x}_0} f(\mathbf{x})| \leq \epsilon.$$

Since this holds for any $\epsilon > 0$, we conclude that $\lim_{n\to\infty} L_n = \lim_{\mathbf{x}\in E, \mathbf{x}\to\mathbf{x}_0} f(\mathbf{x})$. ■

In some contexts (often when dealing with series of functions), the limit function $f(x)$ in the definition of uniform convergence may not be known; then we have the following

Theorem 2.2.2 Cauchy Criterion for Uniform Convergence.

Suppose that $\{f_n(\mathbf{x})\}$ is a sequence of functions defined on a set E. Suppose that for any $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that for any $n, m \geq N$

$$|f_n(\mathbf{x}) - f_m(\mathbf{x})| < \epsilon \textit{ for all } x \in E,$$

then $\{f_n(\mathbf{x})\}$ converges uniformly on E.

Suppose that $\sum_{n=1}^{\infty} a_n(x)$ is a series of functions defined on a set E and that $|a_n(x)| \leq c_n(x)$ for all $x \in E$ and all n. Suppose further that the series $\sum_{n=1}^{\infty} c_n(x)$ converges uniformly on E. Then the series $\sum_{n=1}^{\infty} a_n(x)$ converges uniformly on E.

Remark 2.2.3

Theorem 2.2.1 implies that the answer to question (a) above would be affirmative, if we assume that $f_n(x) \to f(x)$ uniformly over E.

One often cannot apply Theorem 2.2.1 directly on the entire interval on which the question is being asked. For example, we know that the sequence of functions $s_n(x) = \sum_{m=1}^{n} \frac{x^m}{m}$ converges for every $x \in (-1, 1)$, but does not converge uniformly on $(-1, 1)$ (can you provide a proof?). Using the Cauchy Criterion for Uniform Convergence we can also see that, for any $0 < r < 1$, the sequence of $s_n(x) = \sum_{m=1}^{n} \frac{x^m}{m}$ converges uniformly on $[-r, r]$, therefore, making the limit $\sum_{m=1}^{\infty} \frac{x^m}{m}$ a continuous function on $(-r, r)$. Since r is *arbitrary subject to $0 < r < 1$, and any $x_0 \in (-1, 1)$ can be included in some $(-r, r)$ with $0 < r < 1$, this shows that the limit $\sum_{m=1}^{\infty} \frac{x^m}{m}$ a continuous function on $(-1, 1)$. (The limit $\sum_{m=1}^{\infty} \frac{x^m}{m}$ fails to be a uniformly continuous function on $(-1, 1)$, even though each $s_n(x)$ is uniformly continuous on $(-1, 1)$.)*

Exercise 2.2.4 Does the series $\sum_{m=1}^{\infty} mx^m$ converge uniformly over $(-1, 1)$? Does it define a continuous function on $(-1, 1)$?

Exercise 2.2.5 Is it true that if a sequence of continuous functions f_n on E converges uniformly to f on E, then f is bounded on E?

Remark 2.2.6

Although the M-test is an often used tool to check for uniform convergence, there are cases where the M-test doe not apply directly, as in the case of

$$\sum_{k=1}^{\infty} \frac{\sin(kx)}{k},$$

as this series only converges conditionally. Here one needs to check directly whether the Cauchy Criterion for Uniform Convergence holds for the partial sums of this series, namely, we need to examine whether, for any $\epsilon > 0$, we can find N such that for all $n > m \geq N$, we have

$$|\frac{\sin(m+1)x}{m+1} + \cdots + \frac{\sin nx}{n}| < \epsilon \quad \text{for all } x \text{ in a specified set?}$$

This series can be analyzed by applying the summation by parts via Abel's Lemma. Set $s_m(x) = \sum_{k=1}^{m} \sin(kx)$. Then

$$s_m(x) = \frac{\cos\frac{t}{2} - \cos\left(m + \frac{1}{2}\right)t}{2\sin\frac{t}{2}} = \frac{\sin\frac{mt}{2}\sin\frac{(m+1)t}{2}}{\sin\frac{t}{2}}.$$

Note that $s_m(\frac{1}{m})/m \to 2\sin^2\frac{1}{2}$, so $s_m(x)$ does not remain bounded uniformly in m near $x = 0$. However, for any $\pi > \delta > 0$, when x is restricted to $\{x \in [-\pi, \pi] : 0 < \delta < |x|\}$, there exists some $C > 0$ depending on δ such that $|s_m(x)| \leq C$ uniformly in m. Then for such x the Abel summation-by-parts formula gives

$$\begin{aligned}
&|\frac{\sin(m+1)x}{m+1} + \cdots + \frac{\sin nx}{n}| \\
=&\left|\frac{s_{m+1}(x) - s_m(x)}{m+1} + \cdots + \frac{s_n(x) - s_{n-1}(x)}{n}\right| \\
=&\left|s_{m+1}(x)\left(\frac{1}{m+1} - \frac{1}{m+2}\right) + \cdots + s_{n-1}(x)\left(\frac{1}{n-1} - \frac{1}{n}\right)\right| \\
&+\left|\frac{s_n(x)}{n} - \frac{s_m(x)}{m+1}\right| \\
\leq& C\left(\frac{1}{m+1} - \frac{1}{m+2} + \frac{1}{m+2} - \frac{1}{m+3} + \cdots + \frac{1}{n-1} - \frac{1}{n} + \frac{1}{n} + \frac{1}{m+1}\right) \\
\leq&\epsilon,
\end{aligned}$$

provided $n > m \geq N$ for some sufficiently large N depending on ϵ and C (therefore depending on δ). This verifies the Cauchy Criterion for Uniform Convergence on $\{x \in [-\pi, \pi] : 0 < \delta < |x|\}$.

Since this series is 2π periodic, so the result on $[-\pi, \pi]$ can be extended to $\mathbb{R}$ by using the periodic property.

Theorem 2.2.7 Interchange of Limits and Integration.

Suppose that $\{f_n(x)\}$ is a sequence of Riemann-Stieltjes integrable functions with respect to α on (a, b) with $\alpha(b) - \alpha(a)$ finite, and that $f_n(x) \to f(x)$ uniformly over (a, b). Then $f(x)$ is Riemann-Stieltjes integrable functions with respect to α on (a, b) and

$$\int_a^b f(x)\, d\alpha(x) = \lim_{n\to\infty} \int_a^b f_n(x)\, d\alpha(x). \tag{2.2.1}$$

Suppose that $\{a_n(x)\}$ is a sequence of Riemann-Stieltjes integrable functions with respect to α on (a, b) with $\alpha(b) - \alpha(a)$ finite, and that the series $\sum_{n=1}^{\infty} a_n(x)$ converges uniformly over (a, b). Then $\sum_{n=1}^{\infty} a_n(x)$ is Riemann-Stieltjes integrable functions with respect to α on (a, b) and

$$\int_a^b \left(\sum_{n=1}^{\infty} a_n(x)\right) d\alpha(x) = \sum_{n=1}^{\infty} \int_a^b a_n(x)\, d\alpha(x).$$

Proof. For simplicity of presentation we will assume that $\alpha(x)$ is monotone increasing, and set $L = \alpha(b) - \alpha(a)$. For any $\epsilon > 0$, we first use the uniform convergence of $\{f_n(x)\}$ to $f(x)$ on (a, b) to find some N such that for all $n \geq N$ and all $x \in (a, b)$, $|f_n(x) - f(x)| < \epsilon/L$. Then for any $n \geq N$, we find a partition $\mathcal{P}$ of $[a, b]$ such that

$$\int_a^b f_n(x)\, d\alpha(x) - \epsilon \leq L(f_n, \mathcal{P}, \alpha) \leq U(f_n, \mathcal{P}, \alpha) \leq \int_a^b f_n(x)\, d\alpha(x) + \epsilon.$$

Now we find

$$L(f_n, \mathcal{P}, \alpha) - \epsilon \leq L(f, \mathcal{P}, \alpha) \leq U(f, \mathcal{P}, \alpha) \leq U(f_n, \mathcal{P}, \alpha) + \epsilon,$$

which, when combined with the above estimates, gives

$$\int_a^b f_n(x)\, d\alpha(x) - 2\epsilon \leq L(f, \mathcal{P}, \alpha) \leq U(f, \mathcal{P}, \alpha) \leq \int_a^b f_n(x)\, d\alpha(x) + 2\epsilon.$$

This shows that

$$U(f, \mathcal{P}, \alpha) - U(f, \mathcal{P}, \alpha) < 4\epsilon \text{ so } \int_a^b f(x)\, d\alpha(x) \text{ exists;}$$

and that

$$|\int_a^b f(x)\, d\alpha(x) - \int_a^b f_n(x)\, d\alpha(x)| < 2\epsilon.$$

This shows that (2.2.1).

The case for the series follows from the above by applying it to the sequence of partial sums of the series. ■

Remark 2.2.8

The formulation here allows either a or b to be infinite, as long as there is a finite variation $\alpha(b) - \alpha(a)$ over (a, b). For instance, for $a > 0$, the integral $\int_a^\infty f_n(x)\,dx$ could be formulated as $\int_a^\infty f_n(x)x^p\,d\alpha(x)$, with $\alpha(x) = (1-p)^{-1}x^{1-p}$. If for some $p > 1$,

$$f_n(x)x^p \to f(x)x^p \text{ uniformly over } (a, \infty), \tag{2.2.2}$$

we can conclude that

$$\int_a^\infty f_n(x)\,dx = \int_a^\infty f_n(x)x^p\,d\alpha(x) \to \int_a^\infty f(x)x^p\,d\alpha(x) = \int_a^\infty f(x)\,dx.$$

Often one needs some modification to (2.2.2). Suppose that for any finite $b > a$, $f_n(x) \to f(x)$ uniformly over $[a, b]$, and that for some $C > 0, q > 1$, we have

$$|f_n(x)| \le Cx^{-q} \text{ for all } x \in (a, \infty) \text{ and all } n.$$

Then, it follows that $|f(x)| \le Cx^{-q}$ for all $x \in (a, \infty)$, and that for some fixed $p, q > p > 1$,

$$|f_n(x) - f(x)|x^p \le 2Cx^{p-q}$$

So for any $\epsilon > 0$, we can find some $b' > a$ depending on ϵ such that $|f_n(x) - f(x)|x^p \le \epsilon$ for all $x \ge b'$ and all n. But this is not quite the same as (2.2.2). We next discussion an extension to address such scenarios.

Exercise 2.2.9 Identify the pointwise limit $f(x)$ of the sequence $\{\frac{nx}{1+n^2x^2}\}$ for $x \in [0, 1]$. Does it converge uniformly over $x \in [0, 1]$? Does it holds that $\int_0^1 \frac{nx}{1+n^2x^2}\,dx \to \int_0^1 f(x)\,dx$?

Exercise 2.2.10 Identify the pointwise limit $f(x)$ of the sequence $\{\frac{n}{n^2+x^2}\}$ for $x \in [0, \infty)$. Does it converge uniformly over $x \in [0, \infty)$? Does it holds that $\int_0^\infty \frac{n}{n^2+x^2}\,dx \to \int_0^\infty f(x)\,dx$?

2.3 An Extended Criterion for the Interchange of Integration and Limits of a Sequence of Functions

Remark 2.3.1

The conditions in Theorem 2.2.7 often can't be verified on the entity of E, but can be verified after deleting a set of small size, *for example, deleting a small neighbored of one or a finite number of points. If some further* uniform integrability *conditions are assumed, then the conclusions of Theorem 2.2.7 still holds. The discussion below will involve some aspect of* improper integrals.

Theorem 2.3.2 An Extension of Theorem 2.2.7.

Suppose that, for any $c, a < c < b$, $f_n \to f$ uniformly on $[a, c]$, and that each f_n has a convergent integral $\int_a^b f_n(x)\,d\alpha$ (if it is improper at $x = b$).

Furthermore, assume the following uniform integrability of the family $\{f_n\}$:

$$\forall\, \epsilon > 0, \exists\, c, a < c < b, \text{ such that } \forall\, c', c < c' < b, \forall\, n, \left| \int_c^{c'} f_n(x)\, d\alpha \right| < \epsilon. \tag{2.3.1}$$

Then f has a convergent integral $\int_a^b f(x)\, d\alpha$, and

$$\int_a^b f(x)\, d\alpha = \lim_{n\to\infty} \int_a^b f_n(x)\, d\alpha.$$

Proof. First, for any $a' < b', a \le a' < b' < b$, we can apply Theorem 2.2.7 on $[a', b']$ to conclude that

$$\int_{a'}^{b'} f(x)\, d\alpha = \lim_{n\to\infty} \int_{a'}^{b'} f_n(x)\, d\alpha.$$

Next, let c be chosen according to the uniform integrability of the family $\{f_n\}$ in (2.3.1) for a given $\epsilon > 0$. If we set $a' = c$, then for any $c', c < c' < b$, we would get

$$\left| \int_c^{c'} f_n(x)\, d\alpha \right| < \epsilon \text{ for all } n.$$

This then implies that

$$\left| \int_c^{c'} f(x)\, d\alpha \right| = \lim_{n\to\infty} \left| \int_c^{c'} f_n(x)\, d\alpha \right| \le \epsilon,$$

which implies that the integral $\int_a^b f(x)\, d\alpha$ is convergent at $x = b$, and

$$\left| \int_c^b f(x)\, d\alpha \right| \le \epsilon.$$

Finally, using

$$\begin{aligned} &\left| \int_a^b f(x)\, d\alpha - \int_a^b f_n(x)\, d\alpha \right| \\ \le &\left| \int_c^b f(x)\, d\alpha - \int_c^b f_n(x)\, d\alpha \right| + \left| \int_a^c f(x)\, d\alpha - \int_a^c f_n(x)\, d\alpha \right| \\ \le &2\epsilon + \left| \int_a^c f(x)\, d\alpha - \int_a^c f_n(x)\, d\alpha \right| \end{aligned}$$

and applying Theorem 2.2.7 on $[a, c]$, we find some N such that $\left|\int_a^c f(x)\, d\alpha - \int_a^c f_n(x)\, d\alpha\right| < \epsilon$ for all $n \ge N$, which implies

$$\left| \int_a^b f(x)\, d\alpha - \int_a^b f_n(x)\, d\alpha \right| < 3\epsilon$$

for all $n \ge N$, therefore proving the claimed conclusion. ■

Note 2.3.3

What is called uniform integrability *is also called* equi-integrability*. The prefix "equi" here refers uniformity in n, as the notion of* equip-continuity *to be introduced soon.*

Note also that the above formulation allows $b = \infty$, which is a case of an improper integral.

One sufficient condition for (2.3.1) is the existence of an integrable dominating function *g in the sense that*

- *$|f_n(x)| \le g(x)$ for all $\in [a, b), n$ sufficiently large.*
- *$\int_a^b g\, d\alpha$ is convergent.*

In the case that $d\alpha$ gives rise to the usual infinite series, namely, when $\alpha(x) = \sum_n \chi_{\{n \le x\}}(x)$, the condition (2.3.1) takes the form of

$$\forall\, \epsilon > 0, \exists\, N \text{ such that } \forall\, N' > N, \forall\, n, \left| \sum_{m=N}^{N'} f_n(m) \right| < \epsilon.$$

Namely, the tail *part of the summation, $\sum_{m=N}^{\infty} f_n(m)$, can be made small uniformly in n. One sufficient condition for the above is condition of the existence of a similar dominating function: $\exists g(m) \ge 0$ defined for $m \in \mathbb{N}$ such that*

- *$|f_n(m)| \le g(m)$ for all sufficiently large n, m.*
- *$\sum_{m=1}^{\infty} g(m)$ is convergent.*

Examples for the extension. A particular case of an integrable dominating function is a constant function when $\alpha(b)$ is finite. Using this kind of argument one gets

$$\lim_{n\to\infty} \int_0^1 \frac{1}{e^{nx}+1}\, dx = \int_0^1 \lim_{n\to\infty} \frac{1}{e^{nx}+1}\, dx = 0,$$

even though the convergence $\frac{1}{e^{nx}+1} \to 0$ is not uniform over $x \in (0, 1)$.

For the series $\sum_{m=1}^{\infty} \frac{n}{1+nm^2}$, $g(m) = m^{-2}$ is a good dominating function, and we get

$$\lim_{n\to\infty} \sum_{m=1}^{\infty} \frac{n}{1+nm^2} = \sum_{m=1}^{\infty} \lim_{n\to\infty} \frac{n}{1+nm^2} = \sum_{m=1}^{\infty} \frac{1}{m^2}.$$

On the other hand, consider the series $\sum_{m=1}^{\infty} \frac{\chi_{\{m\le n\}}(m)}{n} = \sum_{m=1}^{n} \frac{1}{n} = 1$, the terms $f_n(m) := \frac{\chi_{\{m\le n\}}(m)}{n} \to 0$ uniformly in m as $n \to \infty$, so in this situation

$$\lim_{n\to\infty} \sum_{m=1}^{\infty} \frac{\chi_{\{m\le n\}}(m)}{n} \ne \sum_{m=1}^{\infty} \lim_{n\to\infty} \frac{\chi_{\{m\le n\}}(m)}{n},$$

despite the terms in the series converging to 0 uniformly in m as $n \to \infty$.

Remark 2.3.4

Often we encounter a situation similar to that of Theorem 2.2.7, but we work with a continuum family $\{f(\cdot, y)\}_{y\in I}$ of functions in $\mathcal{R}(\alpha)$ for the parameters y in some metric space I, instead of a sequence of functions. Then the *appropriate modified conclusion should be*

$$\lim_{y\to y_0}\int_a^b f(x,y)\,d\alpha = \int_a^b f(x,y_0)\,d\alpha,$$

and the appropriate modification of the uniform convergence condition should be

$\forall\, \epsilon > 0, \exists\, \delta > 0$ *such that* $|f(x,y)-f(x,y_0)| < \epsilon$ *for all* $d(y,y_0) < \delta, x \in [a,b]$.

A similar modification for (2.3.1) can be formulated.

Exercise 2.3.5 Identify the pointwise limit $f(x)$ of the sequence $\{\frac{n^2}{n^4+x^4}\}$ for $x \in [0,\infty)$. Does it converge uniformly over $x \in [0,\infty)$? Does it holds that $\int_0^\infty \frac{n^2}{n^4+x^4}\,dx \to \int_0^\infty f(x)\,dx$?

Exercise 2.3.6 Does it hold that $\lim_{n\to\infty}\sum_{m=1}^\infty \frac{n}{n^2+m^2} = 0$?

2.4 Interchange of Differentiation and Limit of a Sequence of Functions

Theorem 2.4.1 Interchange of Differentiation and Limit of a Sequence of Functions.

Suppose that $\{f_n(x)\}$ *is a sequence of functions differentiable on* $[a,b]$ *and that* $\{f_n(x_0)\}$ *converges for some* $x_0 \in [a,b]$. *Suppose, in addition, that* $\{f_n'(x)\}$ *converges uniformly on* $[a,b]$. *Then* $\{f_n(x)\}$ *converges uniformly on* $[a,b]$ *to a differentiable function* $f(x)$, *and*

$$f'(x) = \lim_{n\to\infty} f_n'(x).$$

Proof. Let's first give a proof under a stronger assumption: each $f_n'(x)$ is continuous on $[a,b]$, and let's denote $\lim_{n\to\infty} f_n'(x)$ by $g(x)$. Then $g(x)$ is continuous on $[a,b]$, $f_n(x) = f_n(x_0) + \int_{x_0}^x f_n'(t)\,dt$, and Theorem 2.2.7 implies that $\int_{x_0}^x f_n'(t)\,dt \to \int_{x_0}^x g(t)\,dt$, so $\lim_{n\to\infty} f_n(x)$ exists---denote it as $f(x)$ and

$$f(x) = \lim_{n\to\infty} f_n(x_0) + \int_{x_0}^x g(t)\,dt.$$

It follows that

$$f'(x) = g(x) = \lim_{n\to\infty} f_n'(x).$$

Furthermore, the convergence above is uniform over $x \in [a,b]$, as

$$\sup_{x\in[a,b]}\left|\int_{x_0}^x f_n'(t)\,dt - \int_{x_0}^x g(t)\,dt\right| \le \int_a^b |f_n'(t)-g(t)|\,dt \to 0$$

as $n \to \infty$.

We now do a proof in the general case. First, we prove that $\{f_n(x)\}$ converges uniformly over $[a,b]$. Under the assumption of uniform convergence of $\{f_n'(x)\}$ over $[a,b]$, for any $\epsilon > 0$, there exists N such that for $n, m \ge N, x \in [a,b]$, we have

$$|f_n'(x) - f_m'(x)| < \epsilon. \tag{2.4.1}$$

Applying the theorem of the mean to $f_n(x) - f_m(x) - [f_n(x_0) - f_m(x_0)]$, we get

$$f_n(x) - f_m(x) - [f_n(x_0) - f_m(x_0)] = (f_n'(x^*) - f_m'(x^*))\,(x - x_0)$$

for some x^* depending on n, m, x. This leads to

$$|f_n(x) - f_m(x)| \le |f_n(x_0) - f_m(x_0)| + \epsilon(b-a).$$

Since $f_n(x_0) - f_m(x_0) \to 0$ as $n, m \to \infty$, this shows that $\{f_n(x)\}$ satisfies the Cauchy Criterion for Uniform Convergence on $[a, b]$, therefore it converges uniformly to some $f(x)$ on $[a, b]$.

Next for any $x \in (a, b)$ we define $g_n(h) = [f_n(x+h) - f_n(x)]/h$ for $0 < |h| < \delta$ for some $\delta > 0$ (when $x = a$ or b, we restrict h to have appropriate sign). Note that $g_n(h) \to f_n'(x)$ as $h \to 0$ and that $g_n(h) \to [f(x+h) - f(x)]/h$ as $n \to \infty$. We next show that this convergence is uniform for $0 < |h| < \delta$.

We apply the theorem of the mean to $g_n(h) - g_m(h)$ to get

$$g_n(h) - g_m(h) = f_n'(x^*) - f_m'(x^*)$$

for some x^* between x and $x + h$ depending on n, m, x, h. But (2.4.1) holds for $n, m \ge N, x \in [a, b]$, so we get $|g_n(h) - g_m(h)| < \epsilon$ for $n, m \ge N$ uniformly in $0 < |h| < \delta$. This shows that $g_n(h)$ converges to $[f(x+h) - f(x)]/h$ uniformly in $0 < |h| < \delta$ as $n \to \infty$. Now Theorem 2.2.1 applies to $g_n(h)$ to conclude that

$$\lim_{h\to 0} [f(x+h) - f(x)]/h = \lim_{n\to\infty} f_n'(x).$$

Namely, $f'(x)$ exists and equals $\lim_{n\to\infty} f_n'(x)$. ■

Remark 2.4.2

In Theorem 2.4.1, the assumptions are only sufficient, but not necessary conditions. E.g. $f_n(x) = x^n/n$ converges to $f(x) = 0$ uniformly on $I = (-1, 1)$, and $f(x) = 0$ is differentiable on I, yet the convergence of $f_n'(x) = x^{n-1}$ (to $f'(x) = 0$) is not uniform on I (although when restricted to $(-\delta, \delta)$ for any fixed $0 < \delta < 1$, the convergence is uniform.). Another example is $f_n(x) = \sin(nx)/n$, which converges to $f(x) = 0$ uniformly on $\mathbb{R}$, yet $f_n'(x) = \cos(nx)$ would not converge for many values of x (e.g. all rational multiples of π).

Example 2.4.3

Define

$$f(x) = \sum_{k=1}^{\infty} \frac{\cos(kx)}{k^2}.$$

The series converges uniformly for $x \in \mathbb{R}$ so it defines a continuous function on $\mathbb{R}$. To check its differentiability at any x, we need to check whether the differentiated series

$$-\sum_{k=1}^{\infty} \frac{\sin(kx)}{k}$$

converges uniformly in a neighborhood of x, according to Theorem 2.4.1.

We studied this series earlier and showed that for any $\delta, 0 < \delta < \pi$, it converges uniformly when restricted to $\{x \in [-\pi, \pi] : 0 < \delta < |x|\}$. Thus in that region, we do have

$$f'(x) = -\sum_{k=1}^{\infty} \frac{\sin(kx)}{k}.$$

Does $f'(0)$ exist?

Remark 2.4.4

Weierstrass' nowhere differentiable function, which is defined as

$$\sum_{n=0}^{\infty} a^n \cos\left(b^n \pi x\right),\ b \text{ odd},\ 0 < a < 1,\ ab > 1 + \frac{3\pi}{2},$$

is the uniform limit on $\mathbb{R}$ of the infinite series above whose terms are infinitely times differentiable. One can also construct a nowhere differentiable function from the uniform limit of but the building blocks in this construction are not differentiable.

Ideas similar to the Weierstrass's construction show up in later work of J. Nash in constructing C^1 isometric imbedding of a given Riemannian metric, and in more recent work in constructing very rough solutions of Navier-Stokes equations. Roughly speaking, one constructs sufficiently differentiable functions which approximately satisfy the specified equations, but in the limit only a low order regularity is preserved, and the differentiability is lost.

2.5 Metrics on Function Spaces

Another way of viewing Theorem 2.2.1 is that the $C(E)$, the space of continuous functions on E---assuming E to be compact, equipped with the metric

$$\rho_{\sup}(f,g) := \sup_E |f(x) - g(x)|$$

is a *complete* metric space.

On the space $\mathcal{R}(\alpha)$ of Riemann-Stieltjes integrable functions, $\rho_{\sup}(f,g) := \sup_E |f(x) - g(x)|$ is also well defined and becomes a metric on $\mathcal{R}(\alpha)$. Theorem 2.2.7 implies that $\mathcal{R}(\alpha)$ is a complete metric with this metric.

However, in applications we often need to work with another "metric" on $\mathcal{R}(\alpha)$:

$$\rho_{L^1(\alpha)}(f,g) := \int_E |f(x) - g(x)|\, d\alpha.$$

We put a quotation mark on "metric" because it satisfies all the conditions of a metric except for one: $\rho_{L^1(\alpha)}(f,g) = 0$ may not imply that $f = g$ for all $x \in E$. One may use $\rho_{L^1(\alpha)}(f,g) = 0$ to define a relation between two functions $f, g \in \mathcal{R}(\alpha)$, and it's easy to see that this is an equivalence relation. Let's continue to use $\mathcal{R}(\alpha)$ to denote the space of equivalence classes of $\mathcal{R}(\alpha)$ under this equivalence relation.

Question. Is $\mathcal{R}(\alpha)$ a complete metric space under this metric?

The answer turns out to be negative in general, and this turns out a major drawback of Riemann integrable. The main advantage of Lebesgue's integral is that it corrects this deficiency. We will prove that $C([a,b])$ is dense in $\mathcal{R}(\alpha)$ in $\rho_{L^1(\alpha)}$. In Lebesgue's integration theory, it is established that the completion of $C([a,b])$ in $\rho_{L^1(\alpha)}$ is the space of Lebesgue integrable functions.

Example 2.5.1 A metric on $\mathbb{R}^{\mathbb{N}}$.

Note that $d(x, y) := \min\{1, |x-y|\}$ defines a metric on $\mathbb{R}$, which makes $\mathbb{R}$ a complete metric space with bounded diameter---the latter follows because $d(x, y) \le 1$ for all $x, y \in \mathbb{R}$. This metric defines the same topology on $\mathbb{R}$ as the usual Euclidean metric does, namely, a set $U \subset \mathbb{R}$ is open in this metric iff it is open in the usual Euclidean metric.

We can define a metric $\rho(f, g) := \sum_{m=1}^{\infty} \min\{1, |f(m) - g(m)|\}/2^m$ for $f, g : \mathbb{N} \mapsto \mathbb{R}$. A sequence $f_n : \mathbb{N} \mapsto \mathbb{R}$ converges to $f : \mathbb{N} \mapsto \mathbb{R}$ pointwise iff $\rho(f_n, f) \to 0$ as $n \to \infty$.

If we take $f_n(m) = n$ if $m = n$; and $= 0$ if $m \ne n$. Then $f_n(m) \to 0$ pointwise, but not uniformly over $\mathbb{N}$, and $\sum_{m=1}^{\infty} f_n(m) = n \to \infty$ as $n \to \infty$.

Exercise 2.5.2 Prove the assertion that a sequence $f_n : \mathbb{N} \mapsto \mathbb{R}$ converges to $f : \mathbb{N} \mapsto \mathbb{R}$ pointwise iff $\rho(f_n, f) \to 0$ as $n \to \infty$.

Exercise 2.5.3 Prove that the metric space $(\mathbb{R}^{\mathbb{N}}, \rho)$, where ρ is the metric introduced in the above example, is not compact, but for any $M > 0$, the set $\{f \in \mathbb{R}^{\mathbb{N}} : |f(m)| \le M \ \forall m\}$ is a compact set in this metric space.

2.6 Equicontinuous Family of Functions

Definition 2.6.1 Equicontinuous Family.

A family $\mathcal{F}$ of functions f defined on a set E in a metric space (X, d) is said to be **equicontinuous** on E if for every $\epsilon > 0$ there exists some $\delta > 0$ such that

$$|f(x) - f(y)| < \epsilon \text{ whenever } d(x, y) < \delta, x, y \in E, \text{ and } f \in \mathcal{F}. \tag{2.6.1}$$

Remark 2.6.2 On the notion of equicontinuous family of functions.

A fundamental property of the set of real numbers $\mathbb{R}$ *is the Bolzano-Weierstrass Theorem:* any bounded sequence in $\mathbb{R}$ has a convergent subsequence. *One would like to find an appropriate extension of this property on function spaces such as the set* $C(E)$ *of continuous functions on a metric space* E*. Unfortunately, the direct extension does not hold, as demonstrated by the sequence of functions* $\{f_n(x) = \sin nx\}$ *in* $C([0, 2\pi])$.

Using $\int_0^{2\pi} |\sin nx - \sin mx|^2\, dx = 2\pi$ *when* $n \ne m$*, it's clear that* $\{\sin nx\}$ *can't have a subsequence converging uniformly on* $[0, 2\pi]$*. In fact it can't have a subsequence converging pointwise on* $[0, 2\pi]$.

In the situation of the example, what one can directly extend is the following property: there exists a countable dense subset F of $[0, 2\pi]$ and a subsequence $\{f_{n_k}(x)\}$ such that it converges at every $x \in F$. *To be able to say that one can choose a subsequence* $\{f_{n_k}(x)\}$ *such that it converges at every* $x \in [0, 2\pi]$*, one needs to control the behavior of* $\{f_{n_k}(x)\}$ *for* $x \in [0, 2\pi] \setminus F$*. The condition of being equicontinuous is the one what would give us the desired property.*

Heuristically, the equicontinuous condition guarantees that the oscillation of $f_n(x)$ *to be smaller than* $\epsilon > 0$ *for* x *over any neighborhood* V *of suitably* *small diameter, uniformly in n. This allows us to propagate the property*

$$|f_{n_i}(p) - f_{n_j}(p)| < \epsilon \text{ for one point } p \in V$$

to

$$|f_{n_i}(x) - f_{n_j}(x)| < 3\epsilon \text{ for all } x \in V.$$

Using the total boundedness of $E = [a, b]$, this shows that $\{f_{n_i}(x)\}$ is uniformly Cauchy on E.

Definition 2.6.3

A sequence of functions $\{f_n\}$ defined on E is said to be **pointwise bounded** on E if for every $x \in E$ the sequence of scalars $\{f_n(x)\}$ is bounded. $\{f_n\}$ is said to be uniformly bounded on E if there exists $M > 0$ such that

$$|f_n(x)| < M \text{ for all } n \text{ and } x \in E.$$

Proposition 2.6.4 Selecting a Convergent Subsequence.

Suppose that $\{f_n\}$ is a sequence of pointwise bounded functions on a countable *set C. Then it has a subsequence $\{f_{n_k}\}$ such that $\{f_{n_k}(x)\}$ converges for every $x \in C$.*

Proof. Let $\{p_i\}$, $i = 1, \cdots$, be the points of C arranged in a sequence. Then $\{f_n(p_1)\}$ is a bounded sequence, therefore, has a convergent subsequence, say $\{f_{n_{1,k}}(p_1)\}$. Next we pick a convergent subsequence $\{f_{n_{2,k}}(p_2)\}$ from the bounded sequence $\{f_{n_{1,k}}(p_2)\}$. Continuing in this fashion for each p_i, we obtain $\{f_{n_{i,k}}\}$ such that for each i, $\{f_{n_{i,k}}(p_j)\}$ converges for each $j \leq i$ as $k \to \infty$.

Finally, $\{f_{n_{i,i}}\}$ is a subsequence of $\{f_n\}$ which converges at each x_j. ■

Theorem 2.6.5 Ascolli-Arzelà Theorem.

Suppose that K is a compact metric space and that $\{f_n\}$ is a sequence in $C(K)$ pointwise bounded and equicontinuous on K. Then

(a) $\{f_n\}$ is uniformly bounded on K,

(b) $\{f_n\}$ contains a uniformly convergent subsequence.

Proof. For any $\epsilon > 0$, let $\delta > 0$ be such that (2.6.1) holds. Since K is compact, it can be covered by a finite number of sets of diameter $< \delta$, say, $V_1, \cdots, V_m$.

Pick $p_i \in V_i$ for each $i = 1, \cdots, m$. Then for each $i = 1, \cdots, m$, there exists some $M_i > 0$ such that $|f_n(p_i)| \leq M_i$ for all n. Any $x \in V_i$ satisfies

$$|f_n(x) - f_n(p_i)| < \epsilon, \text{ therefore } |f_n(x)| \leq M_i + \epsilon.$$

Let $M = \max_{1 \leq i \leq m} M_i$. Then any $x \in K$ satisfies $|f_n(x)| \leq M + \epsilon$ for all n, which shows (a).

The compactness of K implies that it has a dense countable subset $\{q_i\}$. It then follows from Proposition 2.6.4 that we can pick a subsequence $\{f_{n_k}\}$ such that $\{f_{n_k}(q_i)\}$ converges for each q_i, $i = 1, \cdots$.

Each V_i contains some q_{j_i}. It follows that there exists some N such that $|f_{n_k}(q_{j_i}) - f_{n_l}(q_{j_i})| < \epsilon$ for all $k, l \geq N$ and $i = 1, \cdots, m$.

Using the finite cover $\{V_i\}_{i=1}^m$ of K and (2.6.1) on $\{f_{n_k}\}$, we find that for any $x \in K$, $d(x, q_{j_i}) < \delta$ for some $i = 1, \cdots, m$, therefore,

$$|f_{n_k}(x) - f_{n_l}(x)| \le |f_{n_k}(x) - f_{n_k}(q_{j_i})| + |f_{n_k}(q_{j_i}) - f_{n_l}(q_{j_i})| + |f_{n_l}(q_{j_i}) - f_{n_l}(x)| \le 3\epsilon$$

for all $k, l \ge N$.This show that $\{f_{n_k}\}$ is uniformly Cauchy on K, therefore proving (b). ■

Remark 2.6.6

The Bolzano-Weierstrass/Ascolli-Arzelà property has extensions to function spaces where the convergence is not uniform but in integral norms. We first describe a generalization in the context of the space l^p, defined for $p < \infty$ as the space of infinite sequences $\mathbf{x} := \{\mathbf{x}(k)\}_{k=1}^\infty$ such that

$$\|\mathbf{x}\|_p := \left(\sum_{k=1}^\infty |\mathbf{x}(k)|^p\right)^{1/p} < \infty,$$

we note that, for $p < \infty$, if a sequence $\{\mathbf{x}_n\} \subset l^p$ converges to some $\mathbf{x}_\infty$ in l^p, then for any $\epsilon > 0$, there exists L such that

$$\left(\sum_{k=L}^\infty |\mathbf{x}_\infty(k)|^p\right)^{1/p} < \epsilon,$$

and there exists some N such that for $n \ge N$,

$$||\mathbf{x}_n - \mathbf{x}_\infty||_p = \left(\sum_{k=1}^\infty |\mathbf{x}_n(k) - \mathbf{x}_\infty(k)|^p\right)^{1/p} < \epsilon.$$

This then implies, by Minkowski's inequality, that

$$\left(\sum_{k=L}^\infty |\mathbf{x}_n(k)|^p\right)^{1/p} \le \left(\sum_{k=L}^\infty |\mathbf{x}_n(k) - \mathbf{x}_\infty(k)|^p\right)^{1/p} + \left(\sum_{k=L}^\infty |\mathbf{x}_\infty(k)|^p\right)^{1/p} < 2\epsilon.$$

For the finite number of elements $\{\mathbf{x}_n\}_{n=1}^{N-1}$, we can certainly find some $L' \ge L$ such that, for $n = 1, \ldots, N-1$,

$$\left(\sum_{k=L'}^\infty |\mathbf{x}_n(k)|^p\right)^{1/p} < \epsilon,$$

so

$$\left(\sum_{k=L'}^\infty |\mathbf{x}_n(k)|^p\right)^{1/p} < 2\epsilon, \forall\, n.$$

This turns out to be also a sufficient condition in order to extract a convergent subsequence in l^p. We formulate the condition more precisely in the following.

Definition 2.6.7 Equi-summable Family in l^p.

A family $\{\mathbf{x}_s\}_{s\in I}$ of elements in l^p is said to be **equi-summable** in l^p, if for any $\epsilon > 0$, there exists L such that

$$\left(\sum_{k=L}^{\infty} |\mathbf{x}_s(k)|^p\right)^{1/p} < \epsilon \;\forall\; s \in I. \tag{2.6.2}$$

Theorem 2.6.8 (Sequential) Compactness Condition in l^p.

For $p < \infty$, a sequence $\{\mathbf{x}_n\} \subset l^p$ has a subsequence converging in l^p iff this sequence is equisummable in l^p.

Exercise 2.6.9 Prove that l^p is a complete metric space.

Exercise 2.6.10 Provide a proof of Theorem 2.6.8.

2.7 Weierstrass Theorem

Theorem 2.7.1 Weierstrass Theorem.

Let $-\infty < a < b < \infty$. Every continuous $f : [a,b] \to \mathbb{R}$ can be uniformly approximated by polynomials. In other words, for every continuous $f : [a,b] \to \mathbb{R}$ there is a sequence of polynomials $(p_n(f))_{n\in\mathbb{N}}$ so that

$$\sup_{x\in[a,b]} |p_n(f)(x) - f(x)| \xrightarrow[n\to\infty]{} 0.$$

The proof of Theorem 2.7.1 introduces a very useful idea in analysis: *convolution* and *approximation of identity*.

First, a reduction is done so that we may assume $[a,b] = [0,1]$, and $f(a) = f(0) = 0, f(b) = f(1) = 0$. Next we extend f to be 0 for $x \in \mathbb{R} \setminus [0,1]$, so that it becomes a continuous function on $\mathbb{R}$. Then we define the following convolution of f

$$P_n(x) = \int_{\mathbb{R}} f(s)Q_n(s-x)\,ds$$

where $Q_n(x)$ will be a polynomial , chosen to satisfy the following three properties:

1. $Q_n(x) \geq 0$.
2. $\int_{-1}^{1} Q_n(x)\,dx = 1$ for all n.
3. For any $0 < \delta < 1$, $\int_{-\delta}^{\delta} Q_n(x)\,dx \to 1$ as $n \to \infty$.

Since $Q_n(s-x)$ is a polynomial in x, it is clear that $P_n(x)$ is also a polynomial in x, and using $f = 0$ for $x \leq 0$ or $x \geq 1$, we see that, for $0 \leq x \leq 1$,

$$P_n(x) = \int_{-1+x}^{1+x} f(s)Q_n(s-x)\,ds = \int_{-1}^{1} f(x+t)Q_n(t)\,dt.$$

This indicates that $P_n(x)$ is a weighted average of f, with more weight near $t = 0$ due to item (3) above. We now use items (1)--(3) to see that

$$|f(x) - P_n(x)| = \left|\int_{-1}^{1} (f(x) - f(x+t))\,Q_n(t)\,dt\right|$$

$$\leq \left(\int_{-1}^{\delta} + \int_{\delta}^{1}\right) |f(x) - f(x+t)| Q_n(t)\, dt + \int_{-\delta}^{\delta} |f(x) - f(x+t)| Q_n(t)\, dt$$
$$\leq 2 \max |f| \left(\int_{-1}^{\delta} + \int_{\delta}^{1}\right) Q_n(t)\, dt + \int_{-\delta}^{\delta} |f(x) - f(x+t)| Q_n(t)\, dt.$$

Now for any $\epsilon > 0$, using the uniform continuity of f, we find a $\delta > 0$, such that $|f(x) - f(x+t)| < \epsilon$, for all $t, |t| < \delta$. Then use this δ in the above, and item (3), there exists some N such that for $n \geq N$ we have $2 \max |f| \left(\int_{-1}^{\delta} + \int_{\delta}^{1}\right) Q_n(t)\, dt < \epsilon$, and it follows that

$$|f(x) - P_n(x)| < 2\epsilon.$$

Finally, we can check that, with $Q_n(x) = c_n(1-x^2)^n$, and c_n chosen to satisfy item (2) above, then items (1) and (3) also hold---for item (3), we need to verify that for any $0 < \delta < 1$,

$$\frac{\int_{\delta}^{1}(1-x^2)^n\, dx}{\int_{0}^{\delta}(1-x^2)^n\, dx} \to 0 \text{ as } n \to \infty.$$

Remark 2.7.2

This technique can be extended to the case of continuous functions on a compact subset of $\mathbb{R}^n$---after the integration theory has been extended to that context. In the case of $n = 2$, the P_m's will be polynomials in x and y, not in $z = x + iy$, as one may think that a polynomial in z should be the general extension of a polynomial in x.

Question. If $f \in C^1[a,b]$, can the proof above be adapted to prove that there exists a sequence of polynomials P_m such that

$$P_m \to f \text{ and } P'_m \to f' \text{ uniformly on } [a,b]?$$

2.8 Exercises

1. Prove that, for any finite $a, M > 0$, the set $\{\mathbf{x} := \{\mathbf{x}(k)\}_{k=1}^{\infty} \in l^p : \sum_{k=1}^{\infty} k^a |\mathbf{x}(k)|^p \leq M\}$ is compact in l^p.

2. Suppose that $\{f_n(\mathbf{x})\}$ is a sequence of real valued functions defined on a *compact* set K such that

(a) Each $f_n(x)$ is continuous on K,

(b) $\{f_n(x)\}$ converges pointwise to some *continuous* $f(x)$ on K,

(c) $f_n(x) \geq f_{n+1}(x)$ for all $x \in K$ and $n = 1, 2, \cdots$.

Show that $\{f_n(x)\} \to f(x)$ uniformly on K.

Question: Can the compactness of K be dropped? Where is the continuity of the limit function f in condition (b) used? Can it be dropped?

Solution (A proof using open covering argument). From the assumptions, we know that for any given $\epsilon > 0$, and any $x \in K$, there exist $N = N(x)$ and $\delta_1 = \delta(x) > 0$ such that

$$|f_n(x) - f(x)| < \epsilon \text{ for all } n \geq N,$$

$$|f(t) - f(x)| < \epsilon \text{ for all } t \in B(x, \delta_1),$$

where $B(x, \delta_1)$ stands for the δ neighborhood of x.

Take $n = N(x)$, then, using the continuity of f_n at x, there exists some $\delta_2 > 0$, such that

$$|f_n(t) - f_n(x)| < \epsilon \text{ for all } t \in B(x, \delta_2).$$

Set $\delta = \delta(x) := \min\{\delta_1, \delta_2\}$, then for all $t \in B(x, \delta)$, we have

$$\begin{aligned}|f_n(t) - f(t)| &\leq |f_n(t) - f_n(x)| + |f_n(x) - f(x)| + |f(x) - f(t)| \\ &\leq \epsilon + \epsilon + \epsilon = 3\epsilon.\end{aligned}$$

Furthermore, using the assumption that $f_{n+1}(t) \leq f_n(t)$ for any t, we conclude that $0 \leq f_{n+1}(t) - f(t) \leq f_n(t) - f(t)$, and

$$|f_m(t) - f(t)| \leq |f_n(t) - f(t)| \leq 3\epsilon \text{ for all } m \geq n, t \in B(x, \delta).$$

Now $\{B(x, \delta) : x \in K\}$ forms an open cover of K. Using the compactness of K, we find a finite subcover $\cup_{i=1}^{k} B(x_k, \delta(x_k))$ of K for some $\{x_i\}_{i=1}^{k}$. Set $N = \max_{1 \leq i \leq k} N(x_i)$. Then for any $t \in K$, we have $t \in B(x_i, \delta(x_i))$ for some i, therefore, for all $m \geq N$, we have

$$|f_m(t) - f(t)| \leq 3\epsilon,$$

which concludes our proof.

The above proof can be written in a more compact way. For any $\epsilon > 0$, using the continuity of f_n, f, we know that the set $O_{n,\epsilon} := \{x \in K : f_n(t) - f(t) < \epsilon\}$ is open. Since for any $x \in K$, $f_n(x) - f(x) \to 0$ as $n \to \infty$, we conclude that $x \in O_{n,\epsilon}$ for some n. Thus $\cup_n O_{n,\epsilon}$ is an open over of K. By the compactness of K, there exists some finite subcover. Since $O_{n,\epsilon} \subset O_{m,\epsilon}$ for $m \geq n$, we have in fact $K \subset O_{N,\epsilon}$ for some N, which implies that for all $x \in K$ and all $n \geq N$

$$0 \leq f_n(x) - f(x) \leq f_N(x) - f(x) < \epsilon.$$

3. **Investigate the differentiability of $f(x) = \sum_{k=1}^{\infty} \frac{\cos(kx)}{k^2}$ at $x = 0$.** Note that $f(x)$ is an even function of x, so if $f'(0)$ exists, we would have $f'(0) = 0$. Examine the difference quotient of $f(x)$ at $x = 0$ to see whether it converges to 0.First we have

$$\frac{f(x) - f(0)}{x} = \sum_{k=1}^{\infty} \frac{\cos(kx) - 1}{xk^2} = -2\sum_{k=1}^{\infty} \frac{\sin^2(\frac{kx}{2})}{xk^2} = -2x\sum_{k=1}^{\infty} \left(\frac{\sin(\frac{kx}{2})}{xk}\right)^2.$$

Note that for $x > 0$, $x\sum_{k=1}^{\infty} \left(\frac{\sin(\frac{kx}{2})}{xk}\right)^2$ is a Riemann sum for the improper integral $\int_0^\infty \left(\frac{\sin(y/2)}{y}\right)^2 dy$, so we expect

$$\lim_{x \to 0+} x\sum_{k=1}^{\infty} \left(\frac{\sin(\frac{kx}{2})}{xk}\right)^2 = \int_0^\infty \left(\frac{\sin(y/2)}{y}\right)^2 dy > 0.$$

Theorem 2.2.7 does not directly apply here. One way to justify the assertion above is to use the divide-and-conquer strategy. For any $\epsilon > 0$ given, first we find $L > \epsilon^{-1}$ such that

$$\int_L^\infty \left(\frac{\sin(y/2)}{y}\right)^2 dy < \epsilon.$$

Then, we break the summation into $k \leq x^{-1}L$ and $k > x^{-1}L$. We estimate

$$x \sum_{k>x^{-1}L} \left(\frac{\sin(\frac{kx}{2})}{xk} \right)^2 \leq x^{-1} \sum_{k>x^{-1}L} \frac{1}{k^2} \leq x^{-1} \int_{x^{-1}L}^{\infty} \frac{dy}{y^2} \leq L^{-1} \leq \epsilon.$$

Finally, the summation $x \sum_{k \leq x^{-1}L} \left(\frac{\sin(\frac{kx}{2})}{xk} \right)^2$ is a Riemann sum for the proper Riemann integral $\int_0^L \left(\frac{\sin(y/2)}{y} \right)^2 dy$, so as the step size $x \to 0$, we have

$$\left| x \sum_{k \leq x^{-1}L} \left(\frac{\sin(\frac{kx}{2})}{xk} \right)^2 - \int_0^L \left(\frac{\sin(y/2)}{y} \right)^2 dy \right| < \epsilon.$$

This justifies the limiting process above and concludes that

$$\lim_{x \to 0+} \frac{f(x) - f(0)}{x} = -2 \int_0^{\infty} \left(\frac{\sin(y/2)}{y} \right)^2 dy < 0.$$

Similarly,

$$\lim_{x \to 0-} \frac{f(x) - f(0)}{x} = 2 \int_0^{\infty} \left(\frac{\sin(y/2)}{y} \right)^2 dy > 0.$$

Thus $f'(0)$ does not exist. This proves indirectly that the series for $f'(x)$ can't converge uniformly in a neighborhood of $x = 0$.

Chapter 3

Power Series

3.1 Definition of a Power Series and Its Radius of Convergence

Definition 3.1.1 Definition of a Power Series.

A power series centered at x_0 is a series of the form

$$a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \cdots = \sum_{n=0}^{\infty} a_n(x - x_0)^n$$

for some coefficients $a_0, a_1, a_2, \cdots$.

A basic property of a power series is the following

Proposition 3.1.2 Absolute and Uniform Convergence of a Power Series.

Suppose that the power series $\sum_{n=0}^{\infty} a_n(x - x_0)^n$ converges at some $y \neq x_0$. Then it converges absolutely at any z with $|z - x_0| < |y - x_0|$. In fact, for any $0 < r < |y - x_0|$, the series converges absolutely and uniformly for z such that $|z - x_0| \leq r$.

Proof. Under the assumption that the power series $\sum_{n=0}^{\infty} a_n(x - x_0)^n$ converges at $y \neq x_0$, we know that $a_n(y - x_0)^n \to 0$ as $n \to \infty$. Thus there exists some $C > 0$ such that $|a_n(y - x_0)^n| \leq C$ for all n.

For any z with $|z - x_0| < |y - x_0|$, we can find some $0 < r < |y - x_0|$ such that $|z - x_0| \leq r$. In fact, for any $0 < r < |y - x_0|$, it follows from

$$|a_n(z - x_0)^n| \leq |a_n(y - x_0)^n| \frac{|z - x_0|^n}{|y - x_0|^n} \leq C \left(\frac{r}{|y - x_0|} \right)^n$$

for all z with $|z - x_0| \leq r$ and the comparison test with the geometric series $\sum_{n=0}^{\infty} \left(\frac{r}{|y - x_0|} \right)^n$ that $\sum_{n=0}^{\infty} a_n(z - x_0)^n$ converges absolutely and uniformly for z such that $|z - x_0| \leq r$. Since we can take any such $r < |y - x_0|$, we conclude that $\sum_{n=0}^{\infty} a_n(z - x_0)^n$ converges absolutely at any z with $|z - x_0| < |y - x_0|$. ■

Remark 3.1.3

The proof above shows that the Proposition holds for any complex z such that $|z-x_0| \le r < |y-x_0|$ so the domain of convergence is a round disc centered at x_0. The the disc $D(x_0, R)$ with radius R is the largest disc centered at x_0 in which the power series converges.

Definition 3.1.4 Radius of Convergence of a Power Series.

For any power series $\sum_{n=0}^{\infty} a_n(x-x_0)^n$, we define

$$R = \sup\{|x-x_0| : \sum_{n=0}^{\infty} a_n(x-x_0)^n \text{ converges}\}$$

as the radius of convergence of this power series.

Remark 3.1.5

Note that R could be 0 or ∞, and that, if $0 < R < \infty$, then

(a). *for any $0 < r < R$, the series absolutely and uniformly for x such that $|x-x_0| \le r$,*

thus $\displaystyle\sum_{n=0}^{\infty} a_n(x-x_0)^n$ defines a continuous function for x such that $|x-x_0| < R$;

(b). *for z such that $|x-x_0| > R$, the series diverges at x.*

Theorem 3.1.6 Formula for the Radius of Convergence of a Power Series.

For any power series $\sum_{n=0}^{\infty} a_n(x-x_0)^n$, its radius R of convergence is given by the relation

$$R \limsup_{n\to\infty} |a_n|^{1/n} = 1,$$

where we take $R = 0$ when $\limsup_{n\to\infty} |a_n|^{1/n} = \infty$ and take $R = \infty$ when $\limsup_{n\to\infty} |a_n|^{1/n} = 0$.

Proof. This follows from the root test: when

$$|x-x_0| \limsup_{n\to\infty} |a_n|^{1/n} = \limsup_{n\to\infty} |a_n(x-x_0)^n|^{1/n} < 1,$$

the power series $\sum_{n=0}^{\infty} a_n(x-x_0)^n$ converges absolutely, while when $\limsup_{n\to\infty} |a_n(x-x_0)^n|^{1/n} > 1$, it diverges, for, this condition would imply the existence of a subsequence $n_k \to \infty$ such that $|a_{n_k}(x-x_0)^{n_k}|^{1/n_k} \ge 1$, which would then imply $|a_{n_k}(x-x_0)^{n_k}| \ge 1$, so $\sum_{n=0}^{\infty} a_n(x-x_0)^n$ can't converge at such an x. ∎

Whether $\sum_{n=0}^{\infty} a_n(x-x_0)^n$ converges for x with $|x-x_0| = R$ depends on the particular series, as shown by the following examples

$$\sum_{n=0}^{\infty} x^n, \quad \sum_{n=1}^{\infty} \frac{x^n}{n}, \quad \sum_{n=1}^{\infty} \frac{x^n}{n^2}.$$

All three have 1 as their radius of convergence; but the first series diverges at $x = \pm 1$, the second series converges (conditionally) at $x = -1$ but diverges at $x = 1$, and the third series converges absolutely for all x with $|x| = 1$. Moreover, even though

$\sum_{n=0}^{\infty} x^n$ does not converge at $x = -1$, as $x \to -1+0$, $\sum_{n=0}^{\infty} x^n = (1-x)^{-1} \to \frac{1}{2}$. What about the limiting behavior of $\sum_{n=1}^{\infty} \frac{x^n}{n}$ as $x \to -1+0$? This is answered by the Abel's Theorem to be discussed in the next section.

Since $\limsup_{n\to\infty} |a_n|^{1/n}$ may not always be easy to evaluate, one often needs to provide an estimate for it.

Exercise 3.1.7

1. Prove that $\limsup_{n\to\infty} \sqrt[n]{|a_n|} \le \limsup_{n\to\infty} |\frac{a_{n+1}}{a_n}|$. (One may assume that $a_n \ne 0$ for all n; for, otherwise, the right hand side is ∞.)

2. Prove that $\liminf_{n\to\infty} |\frac{a_{n+1}}{a_n}| \le \liminf_{n\to\infty} \sqrt[n]{|a_n|}$.

3. If $\lim_{n\to\infty} |\frac{a_{n+1}}{a_n}|$ exists, then $\lim_{n\to\infty} \sqrt[n]{|a_n|}$ exists, and equals $\lim_{n\to\infty} |\frac{a_{n+1}}{a_n}|$. (The converse is not true; construct some examples.)

Proposition 3.1.8

The radius R of convergence of the power series $\sum_{n=0}^{\infty} a_n(x-x_0)^n$ satisfies

$$R \ge \left(\limsup_{n\to\infty} |\frac{a_{n+1}}{a_n}|\right)^{-1}, \text{ with equality if } \lim_{n\to\infty} |\frac{a_{n+1}}{a_n}| \text{ exists.}$$

Remark 3.1.9

In computing the radius of convergence of a power series, one often has to make some adaptation. For instance, to determine the radius of convergence of

$$\sum_{m=0}^{\infty} (-1)^m \frac{x^{2m+1}}{(2m+1)!},$$

if one applies the radius of convergence formula directly, then one needs to identify

$$a_n = \begin{cases} \frac{(-1)^m}{(2m+1)!} & \text{if } n = 2m+1 \\ 0 & \text{if } n \text{ is even,} \end{cases}$$

and evaluate $\limsup_{n\to\infty} \sqrt[n]{|a_n|}$, which would involve evaluation of

$$\limsup_{m\to\infty} \{(2m+1)!\}^{\frac{1}{2m+1}},$$

which can be done but requires some work. One could apply the ratio test but needs some adaptation, as $a_n = 0$ for all even n's. One could treat x^{2m+1} as the m-th term, in stead of $(2m+1)$-th term. Then one only needs to make sure that

$$|x|^2 \limsup_{m\to\infty} \frac{(2m+1)!}{(2m+3)!} < 1.$$

This turns out to hold for any x. Thus the radius of convergence here is ∞.

Exercise 3.1.10 Radius of Convergence of a Power Series. Suppose that the radius of convergence of $\sum_{n=0}^{\infty} a_n(x-x_0)^n$ is R. What is the radius of convergence of the series $\sum_{n=0}^{\infty} a_n(x-x_0)^{2n}$ and $\sum_{n=0}^{\infty} a_n^2(x-x_0)^n$ respectively?

3.2 Properties of a Convergent Power Series

Theorem 3.2.1 Abel Theorem for a Power Series.

Suppose that the power series $\sum_{n=0}^\infty a_n(x-x_0)^n$ *converges at* $x=y\neq x_0$*. Then it converges uniformly over the closed interval from* x_0 *to* y*. As a consequence*

$$\lim_{x\to y, |x-x_0|<|y-x_0|} \sum_{n=0}^\infty a_n(x-x_0)^n = \sum_{n=0}^\infty a_n(y-x_0)^n.$$

Proof. Set $c_n = a_n(y-x_0)^n$ and consider the power series $\sum_{n=0}^\infty c_n t^n$ in t. Then it equals the given power series at x with $x-x_0 = t(y-x_0)$ and converges at $t=1$. It suffices to prove that $\sum_{n=0}^\infty c_n t^n$ converges uniformly over $0\le t\le 1$.

Set $s_N = \sum_{n=0}^N c_n$ for $n\ge 0$ and $s_{-1}=0$. Then under our assumption $s_N \to s := \sum_{n=0}^\infty c_n = \sum_{n=0}^\infty a_n(y-x_0)^n$ as $N\to\infty$. For $0\le t<1$, we have

$$\sum_{n=0}^N c_n t^n = \sum_{n=0}^N (s_n - s_{n-1})\, t^n = (1-t)\sum_{n=0}^{N-1} s_n t^n + s_N t^N.$$

Sending $N\to\infty$ and using $(1-t)\sum_{n=0}^\infty t^n = 1$ gives us

$$\sum_{n=0}^\infty c_n t^n = (1-t)\sum_{n=0}^\infty s_n t^n = (1-t)\sum_{n=0}^\infty (s_n - s)\, t^n + s.$$

For any $\epsilon>0$, let N be such that for $n>N$, we have $|s_n - s|<\epsilon$. Then

$$\begin{aligned}
|\sum_{n=0}^\infty c_n t^n - s| &\le |1-t|\sum_{n=0}^\infty |s_n - s||t|^n \\
&= |1-t|\left(\sum_{n=0}^N |s_n - s||t|^n + \sum_{n=N+1}^\infty |s_n-s||t|^n\right) \\
&\le |1-t|\left(\sum_{n=0}^N |s_n - s||t|^n + \epsilon\sum_{n=N+1}^\infty |t|^n\right) \\
&\le |1-t|\left(\sum_{n=0}^N |s_n - s||t|^n\right) + \epsilon\frac{|t|^{N+1}|1-t|}{1-|t|}.
\end{aligned}$$

If we take real t such that $0\le t<1$, then $\frac{|t|^{N+1}|1-t|}{1-|t|}\le 1$, and we can find some $\delta>0$ such that when $|1-t|<\delta$, we have $|1-t|\left(\sum_{n=0}^N |s_n-s||t|^n\right)<\epsilon$. This shows that $\sum_{n=0}^\infty c_n t^n \to s$ as $t\to 1-$. Note that the above argument works even for complex t as long as we restrict t to satisfy $\frac{|1-t|}{1-|t|}\le C$ for some $C>0$. This is the case as long as $t\to 1$ from within the unit disc in a *non-tangential* way. ■

Remark 3.2.2

Abel's Theorem is a form of interchange of limits. Suppose that the series $\sum_{n=0}^\infty x^n$ *converges at one end, say at* $x=R$*, then the Theorem implies that* *for* $0 \le x \le R$

$$\sum_{n=1}^{\infty} a_n x^n = \lim_{N\to\infty} s_N(x)$$

is a continuous function of $x \in [0, R]$*. As a consequence,*

$$\sum_{n=1}^{\infty} a_n R^n = \lim_{x\to R-} \sum_{n=1}^{\infty} a_n x^n.$$

This can also be written as

$$\lim_{N\to\infty} s_N(R) = \lim_{N\to\infty} \lim_{x\to R-} s_N(x) = \lim_{x\to R-} \lim_{N\to\infty} s_N(x) = \lim_{x\to R-} \sum_{n=1}^{\infty} a_n x^n.$$

Theorem 3.2.3 Term-wise Integration of a Convergent Power Series.

Suppose that the radius R *of convergence of the power series* $\sum_{n=0}^{\infty} a_n(x-x_0)^n$ *is non zero. Then for any* $0 < t < R$*,*

$$\int_{x_0}^{x_0+t} \left(\sum_{n=0}^{\infty} a_n(x-x_0)^n \right) dx = \sum_{n=0}^{\infty} \int_{x_0}^{x_0+t} a_n(x-x_0)^n \, dx = \sum_{n=0}^{\infty} \frac{a_n t^{n+1}}{n+1}.$$

Proof. This follows simply using the knowledge that the power series converges uniformly for x such that $x_0 \le x \le x_0 + t$. ■

Example 3.2.4 The series $\sum_{n=0}^{\infty}(-x^2)^n$.

It arises from the geometric series replacing x by $-x^2$ when $|x| < 1$:

$$\frac{1}{1+x^2} = \sum_{n=0}^{\infty}(-x^2)^n = \sum_{n=0}^{\infty}(-1)^n x^{2n} = 1 - x^2 + x^4 - x^6 + \cdots$$

Its radius of convergence is 1. Thus for any t, $|t| < 1$, we have

$$\int_0^t \frac{1}{1+x^2} \, dx = \sum_{n=0}^{\infty} \int_0^t (-1)^n x^{2n} \, dx = \sum_{n=0}^{\infty} \frac{(-1)^n t^{2n+1}}{2n+1}.$$

Using calculus knowledge that $\int_0^t \frac{1}{1+x^2} \, dx = \arctan t$, and the series on the right converges at $t = 1$, Theorem 3.2.1 implies that

$$\frac{\pi}{4} = \lim_{t\to 1-0} \arctan t = \sum_{n=0}^{\infty} \lim_{t\to 1-0} \frac{(-1)^n t^{2n+1}}{2n+1} = \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1}.$$

Theorem 3.2.5 Infinite Differentiability of a Convergent Power Series.

Suppose that the radius R *of convergence of the power series* $\sum_{n=0}^{\infty} a_n(x - x_0)^n$ *is non zero. Then the power series defines an infinitely many times* *differentiable function $f(x)$ of x for $|x-x_0|<R$,*

$$f'(x)=\sum_{n=1}^{\infty} na_n(x-x_0)^{n-1} \quad \textit{for } |x-x_0|<R, \tag{3.2.1}$$

and for all $k\geq 1$

$$f^{(k)}(x)=\sum_{n=k}^{\infty} n(n-1)\cdots(n-k+1)a_n(x-x_0)^{n-k} \quad \textit{for } |x-x_0|<R,$$

and

$$f^{(k)}(x_0)=k!a_k.$$

Proof. One simply notes that the series $\sum_{n=1}^{\infty} na_n(x-x_0)^{n-1}$ converges iff the series $\sum_{n=1}^{\infty} na_n(x-x_0)^{n}$ converges, so these two series have the same radii of convergence. But the radius of convergence of the latter is given by

$$\limsup_{n\to\infty}|na_n|^{1/n}=\limsup_{n\to\infty} n^{1/n}|a_n|^{1/n}=\limsup_{n\to\infty}|a_n|^{1/n}$$

using $\lim_{n\to\infty} n^{1/n}=1$. As a result, for any $0<r<R$, the series $\sum_{n=1}^{\infty} na_n(x-x_0)^{n-1}$ converges uniformly over x such that $|x-x_0|\leq r$, and we can can apply the theorem on a sequence of functions whose derivative sequence converges uniformly to conclude that (3.2.1) holds for $|x-x_0|\leq r$. Since this holds for any $0<r<R$, we conclude that (3.2.1) holds for $|x-x_0|<R$. The rest cases for $k>1$ follow by repeating this procedure. ■

Remark 3.2.6

The previous Theorem reveals that a convergent power series in $x-x_0$ is the Taylor series of the power series at x_0. This raises the question: whether any infinitely differentiable function can be represented near any point x_0 in its domain as a convergent power series in $x-x_0$?

If this holds true for such a function $f(x)$, the power series must be the Taylor series of f at x_0. This raises a related question: does the Taylor series at any x_0 in its domain of an infinitely differentiable function always converge? It turns out that, even if this Taylor series converges, the convergent power series may not be equal to f

If $f(x)$ is infinitely differentiable near x_0, then the Taylor expansion with remainder term gives us, for any N,

$$f(x)=\sum_{n=0}^{N}\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n+\frac{f^{(N+1)}(c)}{(n+1)!}(x-x_0)^{N+1}$$

for some c between x_0 and x. So our questions above boil down to whether $\frac{f^{(N+1)}(c)}{(n+1)!}(x-x_0)^{N+1}\to 0$ as $N\to\infty$. This clearly requires some control on $|f^{(N+1)}(c)|$.

Exercise 3.2.7 An infinitely differentiable function which does not equal its Taylor series at some point. Consider

$$f(x)=\begin{cases} e^{-\frac{1}{x^2}} & \text{if } x\neq 0;\\ 0 & \text{if } x=0.\end{cases}$$

Verify that this $f(x)$ is infinitely differentiable and $f^{(k)}(0) = 0$ for all k. Note that its Taylor series at 0 is identically 0 so can't equal $f(x)$ for $x \neq 0$. We remark that if we choose any $x_0 \neq 0$, then it turns out that $f(x)$ equals its Taylor series at x_0 in a neighborhood of x_0.

Theorem 3.2.8 A Convergent Power Series' Expansion at a Different Center.

Suppose that the radius R of convergence of the power series $\sum_{n=0}^{\infty} a_n(x-x_0)^n$ is non zero and x_1 satisfies $|x_1 - x_0| < R$. Let $r = R - |x_1 - x_0|$. Then the following holds

$$\sum_{n=0}^{\infty} a_n(x-x_0)^n = \sum_{k=0}^{\infty} b_k(x-x_1)^k$$

at least for all x such that $|x - x_1| < r = R - |x_1 - x_0|$, where

$$b_k = \sum_{n=k}^{\infty} \binom{n}{k} a_n(x_1 - x_0)^{n-k}.$$

Proof. We use the binomial expansion

$$(x-x_0)^n = (x - x_1 + x_1 - x_0)^n = \sum_{k=0}^{n} \binom{n}{k} (x-x_1)^k (x_1 - x_0)^{n-k}$$

to write the power series $\sum_{n=0}^{\infty} a_n(x-x_0)^n$ as a doubly indexed series. We show that this doubly indexed series converges absolutely for any x such that $|x - x_1| < r = R - |x_1 - x_0|$. As a result, we can exchange the order of summation to obtain

$$\sum_{n=0}^{\infty}\sum_{k=0}^{n} a_n \binom{n}{k} (x-x_1)^k (x_1-x_0)^{n-k} = \sum_{k=0}^{\infty}\sum_{n=k}^{\infty} a_n \binom{n}{k} (x_1-x_0)^{n-k}(x-x_1)^k,$$

which is the desired $\sum_{k=0}^{\infty} b_k(x-x_1)^k$.

The absolute convergence of the doubly indexed series is seen by noting that

$$\sum_{n=0}^{\infty}\sum_{k=0}^{n} |a_n| \binom{n}{k} |x-x_1|^k |x_1-x_0|^{n-k} = \sum_{n=0}^{\infty} |a_n|(|x-x_1| + |x_1-x_0|)^n$$

and that $|x-x_1| + |x_1 - x_0| < R$ under our assumption $|x - x_1| < r = R - |x_1 - x_0|$, and the knowledge that the series $\sum_{n=0}^{\infty} |a_n| s^n$ converges for any $s < R$. ■

Remark 3.2.9

It is possible for the radius of convergence of the power series $\sum_{k=0}^{\infty} b_k(x-x_1)^k$ to be greater than $R - |x_1 - x_0|$, as seen in the case of $\sum_{n=0}^{\infty} x^n$ and $x_1 = -\frac{1}{2}$, where, instead of computing the Taylor series of $(1-x)^{-1}$ centered at $x_1 = -\frac{1}{2}$, we can use a geometric series to expand $\sum_{n=0}^{\infty} x^n = (1-x)^{-1}$ as

$$(1-x)^{-1} = \frac{1}{\frac{3}{2} - (x + \frac{1}{2})} = \frac{2}{3}\frac{1}{1 - \frac{x+\frac{1}{2}}{\frac{3}{2}}} = \frac{2}{3}\left(\sum_{n=0}^{\infty}\left(\frac{x+\frac{1}{2}}{\frac{3}{2}}\right)^n\right),$$

which converges as long as $|x + \frac{1}{2}| < \frac{3}{2}$.

Theorem 3.2.10 A Non-zero Convergent Power Series Has Isolated Zeroes.

Suppose that the radius R of convergence of the power series $\sum_{n=0}^{\infty} a_n(x-x_0)^n$ is non zero, and that it has a sequence of distinct zeroes approaching some x_1 with $|x_1 - x_0| < R$, then it must be identically 0 for all x with $|x - x_0| < R$.

Proof. We first represent the series as a power series in $x-x_1$ for $|x-x_1| < R-|x_1-x_0|$ by $\sum_{k=0}^{\infty} b_k(x - x_1)^k$. We argue by contradiction. Suppose that the series is not identically 0, then there is a smallest integer $k = m$ such that $b_k \neq 0$. It follows that

$$\sum_{k=0}^{\infty} b_k(x - x_1)^k = \sum_{k=m}^{\infty} b_k(x - x_1)^k = (x - x_1)^m \sum_{k=m}^{\infty} b_k(x - x_1)^{k-m}.$$

The series $\sum_{k=m}^{\infty} b_k(x - x_1)^{k-m}$ has the same non-zero radius of convergence, $R - |x_1 - x_0|$, as that for $\sum_{k=0}^{\infty} b_k(x - x_1)^k$, so defines a continuous function $g(x)$ in a neighborhood of x_1. Since $g(x_1) = b_m \neq 0$, we conclude that $g(x) \neq 0$ in a neighborhood of x_1. But this would mean that $\sum_{n=0}^{\infty} a_n(x - x_0)^n = (x - x_1)^m g(x)$ has no zero beside x_1 in this neighborhood, contradicting our assumption that a sequence of zeroes of $\sum_{n=0}^{\infty} a_n(x - x_0)^n$ approach x_1. This shows that under our assumption the power series must have all its $b_k = 0$, and that it equals 0 for all x such that $|x - x_1| < R - |x_1 - x_0|$.

The above argument applies to any z with $|z - x_0| < R$ when z is a limit point of the set of zeroes of the power series. Let Z be the set of such points in $|z - x_0| < R$. Then Z is a non-empty closed subset of the disc $|x - x_0| < R$. But the above argument shows that Z is also open. This then implies that Z must consists of all points of the disc $|x - x_0| < R$, therefore, the power series equals 0 in this entire disc. ■

Remark 3.2.11

In the above theorem, it is important that the limit point referred to is in the disc of convergence of the power series. The function $f(x) = (x + 1)\sin\frac{1}{x+1}$ has a convergent power series expansion at $x = 0$ with radius of convergence equal to 1, and $f(x_n) = 0$ for $x_n = -1 + \frac{1}{n\pi} \to -1$, yet $f(x) \not\equiv 0$.

Theorem 3.2.12 Product of Two Power Series.

*Suppose that $f(x) = \sum_{n=0}^{\infty} a_n x^n$ and $g(x) = \sum_{n=0}^{\infty} b_n x^n$ are two power series converging for $|x| < R$. Then $f(x)g(x)$ has a power series expansion $\sum_{n=0}^{\infty} c_n x^n$ in x converging for $|x| < R$, and $c_n = a_0 b_n + a_1 b_{n-1} + \cdots + a_n b_0$ is the **Cauchy product** of the sequence $\{a_n\}$ and $\{b_n\}$.*

Proof. For any x with $|x| < R$, we know that both $f(x) = \sum_{n=1}^{\infty} a_n x^n$ and $g(x) = \sum_{n=1}^{\infty} b_n x^n$ converge absolutely, so $f(x)g(x)$ equals the Cauchy product given as

$$\sum_{n=0}^{\infty}\left(\sum_{m=0}^{n} a_m b_{n-m}\right) x^n.$$

■

Combing this Theorem and Abel's Theorem gives us

Corollary 3.2.13

Suppose that $\sum_{n=0}^{\infty} a_n$, $\sum_{n=0}^{\infty} b_n$, *and* $\sum_{n=0}^{\infty} b_n$ *all converge, where* $c_n = \sum_{m=0}^{n} a_m b_{n-m}$ *is the Cauchy product of sequence* $\{a_n\}$ *and* $\{b_n\}$. *Then*

$$\left(\sum_{n=0}^{\infty} a_n\right)\left(\sum_{n=0}^{\infty} b_n\right) = \sum_{n=0}^{\infty} c_n.$$

Proof. Set $f(x) = \sum_{n=0}^{\infty} a_n x^n$ and $g(x) = \sum_{n=0}^{\infty} b_n x^n$. Then the two power series converge at $x = 1$. Thus $f(x)g(x) = \sum_{n=0}^{\infty} c_n x^n$ for all x with $|x| < 1$. Abel Theorem is applicable to all three power series here for $x = 1$, so

$$\left(\sum_{n=0}^{\infty} a_n\right)\left(\sum_{n=0}^{\infty} b_n\right) = \lim_{x\to 1-}\left(\sum_{n=0}^{\infty} a_n x^n\right)\left(\sum_{n=0}^{\infty} b_n x^n\right) = \lim_{x\to 1-}\sum_{n=0}^{\infty} c_n x^n = \sum_{n=0}^{\infty} c_n.$$

■

Remark 3.2.14

The condition that all three series converge can't be dropped. For the case that $a_n = b_n = \frac{(-1)^n}{n+1}$, *it turns out that the series of the Cauchy product* $\sum_{n=0}^{\infty} c_n$ *does not converge.*

Remark 3.2.15

It can be proved that the quotient of two convergent power series centered at some x_0 *has a convergent power series expansion centered at* x_0 *in some neighborhood of* x_0, *provided that the denominator does not vanish at* x_0. *Similarly it can be proved that if* $f(x)$ *has a convergent power series expansion centered at* x_0 *with* $y_0 = f(x_0)$, *and* $g(y)$ *has a convergent power series expansion centered at* y_0, *then the composition* $g \circ f(x)$ *has a convergent power series expansion centered at* x_0 .

3.3 Exponential Functions

The discussion of the previous section applies to general power series. It is more interesting to discuss some special power series that arise from important applications or have special properties. The exponential and logarithmic functions are two families of such functions.

In most elementary calculus treatment, the definition of the exponential functions a^x and proof of their properties use some hand waving at some points; similarly, the definition and properties of the trigonometric functions rely on some geometric arguments, instead of purely analytical ones. It should be a rewarding experience to review such a treatment and pinpoint such places.

The exponential functions arise most naturally as solutions of the ODE $y'(x) = ry(x)$. In looking for a solution of the form $y = \sum_{n=0}^{\infty} c_n x^n$, one finds

$$y'(x) = \sum_{n=0}^{\infty} c_n n x^{n-1} = ry(x) = r\sum_{n=0}^{\infty} c_n x^n,$$

from which one concludes that $c_n n = rc_{n-1}$. Then by induction one gets

$$c_n = c_0 \frac{r^n}{n!}.$$

Thus $y = c_0 \sum_{n=0}^{\infty} \frac{r^n x^n}{n!}$ should be a solution with $y(0) = c_0$. To justify the argument, one checks that this power series has its radius of convergence equal to ∞, so all the derivations are justified.

How was the solution $c_0 e^{rx}$ introduced in calculus? And how does this power series solution relate to $c_0 e^{rx}$? Is this the only solution with $y(0) = c_0$? The answer lies in developing properties of this solution. Denoting $E(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!}$. The key properties are

$$\begin{aligned} E'(x) &= E(x), \ \forall x, \\ E(x+y) &= E(x)E(y), \ \forall x, y, \\ E(0) &= 1. \end{aligned}$$

It then follows that $E(-x)E(x) = E(0) = 1$ for any x, so $E(x) \neq 0$ for any x (even complex valued). When x takes real values, $E(x)$ also takes real values by construction. For $x > 0$, the power series for $E(x)$ shows that $E(x) > 0$. One then shows using $E(-x)E(x) = E(0) = 1$ that $E(x) > 0$ for all real $x < 0$. Then the property $E'(x) = E(x) > 0$ shows that $E(x)$ is monotone increasing for real valued x.

At this point, there is a well defined inverse function of $E(x)$ for $x \in \mathbb{R}$. Call it $\ln y$ for $y > 0$. Then $E(\ln y) = y$ and $\ln E(x) = x$. It remains to establish that $\ln y$ is defined for all $y > 0$. As a consequence of $E(x + y) = E(x)E(y)$, we will have $\ln(uv) = \ln u + \ln v$ for $u, v > 0$.

Based on the properties of $E(x)$, one establishes that for any rational $x = \frac{p}{q}$,

$$E(x) = [E(\frac{1}{q})]^p = [E(1)]^{\frac{p}{q}} = [E(1)]^x.$$

If one can establish that the function a^x is well defined for any real $a > 0$ and any real x and that it is a continuous function for $x \in \mathbb{R}$, then one can use the continuity to show that the above equality holds for all x. However it is not a trivial task to define a^x for any real $a > 0$ and any real x and prove that it is a continuous function of x.

Recall that there is an arithmetic procedure for computing ab and a^m only when a, b are rational numbers and m is an integer, that the definition of $a^{1/n}$ for any positive real (even a rational number) a and positive integer n requires a limiting process and completeness of $\mathbb{R}$. Once this is defined, one can use the continuity and monotonicity of the power function $x \mapsto x^k$ for any positive integer k and positive real x to define $a^{m/n}$ for any positive real a and positive integers m, n, namely, a^r for any positive real a and positive rational r. Additional work is needed to define a^x for any positive real a and any real x, as was done in Exercise 6 of Chapter 1 of Rudin's text for the case of $a > 1$.

Using properties of power series, the definition and properties of $E(x)$ can be developed in a routine way, which is how Rudin develops this material. Rudin also sketches an argument to show that the treatment following Exercise 6 of Chapter 1 produces the same function as that using $E(x)$.

3.4 The Trigonometric Functions

The trigonometric functions arise from geometric considerations. It was Euler who discovered the relation between the trigonometric functions and the exponential function as encoded in the Euler formula

$$E(ix) = \cos(x) + i\sin(x).$$

$E(ix)$ is defined in terms of the power series expansion

$$\sum_{n=0}^{\infty} \frac{(ix)^n}{n!}.$$

For real valued x, splitting the above series into its real and imaginary parts, we obtain

$$\cos x = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{(2k)!}, \quad \sin x = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)!}.$$

But the series converges for all complex valued x, so we also define $\cos x$ and $\sin x$ for complex valued x using the above series.

Note that the definitions for these trigonometric functions are in purely analytic means, thus these trigonometric functions as defined this way do not directly have relations with the ones defined geometrically. Rudin's development follows this approach and does not use any of the properties as given by the geometric approach. For instance, he sets out to prove that $E(x)$ has a purely imaginary period, labeled as $2\pi i$, and use this to show that both $\cos x$ and $\sin x$ also have $2\pi i$ as their period; but the 2π here does not have a direct relation with the angle interpretation in geometry. This kind of treatment is fine for a rigorous development of calculus, but should not be taken as a discouragement from relating to the geometric approach. In fact it is much more productive to fully use the geometric interpretation; one just needs to be ware of the places where a certain geometric properties play a crucial role and how those arguments can be replaced by purely analytical ones.

3.5 Exercises

1. **Binomial Power Series Expansion.** For any real number α and nonnegative integer n, define $\binom{\alpha}{n} = \alpha(\alpha-1)\cdots(\alpha-n+1)/n!$. Show that the radius of convergence of the power series $\sum_{n=0}^{\infty}\binom{\alpha}{n}x^n$ equals 1. Then show that $(1+x)^\alpha = \sum_{n=0}^{\infty}\binom{\alpha}{n}x^n$ for $|x|<1$.

Hint. Estimate the remainder term in the Taylor expansion of $(1+x)^\alpha$.

2. Suppose that the power series $\sum_{n=0}^{\infty} a_n x^n$ has radius of convergence equal to 1, that each $a_n \geq 0$ and that $\sum_{n=0}^{\infty} a_n = \infty$. Show that $\lim_{x\to 1-}\sum_{n=0}^{\infty} a_n x^n = \infty$.

3. **Composition of Convergent Power Series.** Suppose that the power series $f(x) = \sum_{n=0}^{\infty} a_n x^n$ has a positive radius of convergence, and $g(y) = \sum_{n=0}^{\infty} b_n (y-a_0)^n$ also has a positive radius of convergence R. Prove that for all x such that $\sum_{n=1}^{\infty}|a_n x^n| < R$, the composite function $g\circ f(x)$ has a convergent power series expansion of the form $\sum_{n=0}^{\infty} c_n x^n$, where $c_0 = b_0$ and $c_n = \sum_{k=0}^{n} b_k a_n(k)$ for positive integer n, with $a_n(k)$ defined from the relation $(\sum_{k=1}^{\infty} a_k x^k)^n = \sum_{k=1}^{\infty} a_k(n) x^k$.

Hint. Substitute $y = f(x)$ into the power series expansion of $g(y)$ in $y - a_0$ and justify the interchange of order of summation.

Chapter 4

Fourier Series

Fourier series arise naturally when constructing solutions of certain initial-boundary value problems of partial differential equations (PDEs). For example, they arise from studying the initial-boundary value problem for the heat equation

$$\begin{aligned} u_t(x,t) - u_{xx}(x,t) &= 0 && 0 < x < l, t > 0, \\ u(0,t) = u(l,t) &= 0 && t > 0, \\ u(x,0) &= g(x) && 0 < x < l, \end{aligned}$$

where the initial data $g(x)$ is a given continuous function on $[0,l]$, and traditionally we would like the solution $u(x,t)$ to be twice continuously differentiable in x, once continuously differentiable in t in the domain $(0,l) \times (0,\infty)$, and continuous on $[0,l] \times [0,\infty)$.

There is an elementary procedure of looking for separable particular solutions $u(x,t)$ of the form $X(x)T(t)$, which solves the homogeneous heat equation and the homogeneous boundary conditions. The result is that for any $n \in \mathbb{N}$,

$$u_n(x,t) := \sin\left(\frac{n\pi x}{l}\right) e^{-\left(\frac{n\pi}{l}\right)^2 t}$$

is such a solution. Since we are so far dealing with linear homogeneous equations, any linear combination of solutions is still a solution, so

$$\sum_{n \in \text{a finite set}} c_n \sin\left(\frac{n\pi x}{l}\right) e^{-\left(\frac{n\pi}{l}\right)^2 t}$$

also satisfies the same equations. What remains is whether one can choose the c_n's so that this solution at $t = 0$ gives rise to the prescribed initial data $g(x)$.

For that purpose, first we need to form an infinite sum and demand that

$$\sum_{n=1}^{\infty} c_n \sin\left(\frac{n\pi x}{l}\right) = g(x) \quad \text{on } (0,l) \text{ in an appropriate sense.} \tag{4.0.1}$$

But we also need to make sense of the infinite series as a continuously differentiable solution.

The expansion (4.0.1) is a version of the Fourier series expansion. It turns out that we must choose c_n such that

$$c_n = \frac{2}{l} \int_0^l g(x) \sin\left(\frac{n\pi x}{l}\right) \, dx. \tag{4.0.2}$$

The key properties that lead to the formula (4.0.2) for c_n and many other properties of the series are the **orthogonality relations** of the terms over the interval $[0, l]$ stated below:

$$\int_0^l \sin\left(\frac{n\pi x}{l}\right) \sin\left(\frac{m\pi x}{l}\right) dx = \begin{cases} 0 & \text{if } n \neq m, \\ \frac{l}{2} & \text{if } n = m. \end{cases}$$

We multiply $\sin\left(\frac{m\pi x}{l}\right)$ to both sides of (4.0.1) and integrate both sides over $[0, l]$. If we can justify the interchange of summation and integration, then the orthogonality relations above would lead to (4.0.2).

If the most elementary notion of pointwise convergence is used in (4.0.1), then it is not so easy to justify the interchange of summation and integration. It turns out that a more useful notion of convergence in this context is that of **mean square convergence** defined as

$$\lim_{N\to\infty} \int_0^l |g(x) - \sum_{n=1}^{N} c_n \sin\left(\frac{n\pi x}{l}\right)|^2\, dx \to 0.$$

4.1 General Orthogonal Expansion

The notion of mean square convergence makes senes in a general inner product space.

Definition 4.1.1 Inner Product Space.

A vector space V over the reals $\mathbb{R}$ is called an inner product space if there is a function $(x, y) \in V \times V \mapsto (x, y) \in \mathbb{R}$ such that

1. $(a_1x_1 + a_2x_2, y) = a_1(x_1, y) + a_2(x_2, y)$ for any $x_1, x_2, y \in V$ and any $a_1, a_2 \in \mathbb{R}$;
2. $(x, y) = (y, x)$ for any $x, y \in V$;
3. $(x, x) \geq 0$ for any $x \in V$ and equals 0 iff $x = 0$.

Note that the first two properties imply

$$(x, a_1y_1 + a_2y_2) = a_1(x, y_1) + a_2(x, y_2) \text{ for any } x, y_1, y_2 \in V.$$

Due to this and the first property, we say an inner product on a vector over the reals is **bilinear**.

The space of real valued continuous function on a finite interval $[a, b]$, $\mathcal{C}[a, b]$, has a natural inner product: $(f, g) := \int_a^b f(x)g(x)\, dx$.

Definition 4.1.2 Orthogonal Relation.

Two vectors x, y in an inner product space V are said to be orthogonal to each other if $(x, y) = 0$.

Note that, since $(y, x) = (x, y)$, it follows that if $(x, y) = 0$, then $(y, x) = 0$. So the orthogonal relation is symmetric in x and y.

In the context of $\mathcal{C}[a, b]$, two functions $f, g \in \mathcal{C}[a, b]$ are orthogonal in $\mathcal{C}[a, b]$ if $\int_a^b f(x)g(x)\, dx = 0$. Note that it is important to specify the interval of integration.

The orthogonality relation stated earlier says that, when $n \neq m$, the functions $\sin\left(\frac{n\pi x}{l}\right), \sin\left(\frac{m\pi x}{l}\right)$ are orthogonal on $[0, l]$. But these two functions may not be orthogonal on a different interval such as $[0, l/2]$.

Proposition 4.1.3 Basic Properties of an Inner Product Space.

*Suppose that V is an inner product space over the reals $\mathbb{R}$. Then the **Cauchy-Schwarz inequality** holds:*

$$|(x, y)| \leq \sqrt{(x, x)}\sqrt{(y, y)} \text{ for all } x, y \in V.$$

$\|x\| := \sqrt{(x, x)}$ *defines a norm on V:*

1. *$\|x\| \geq 0$ for all $x \in V$ and equals 0 iff $x = 0$ in V;*
2. *$\|ax\| = |a|\|x\|$ for all $x \in V$ and real $a \in \mathbb{R}$;*
3. *$\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in V$ (Triangle Inequality).*

When $(x, y) = 0$, namely, when x is orthogonal to y in V, we also have the Pythagorean relation:

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2.$$

Proof. For any $x, y \in V$, the function $t \mapsto (x + ty, x + ty)$ in t is a quadratic function in t when $y \neq 0$ due to the bilinear property of the inner product, and is nonnegative. Its minimum is attained at $t = -(x, y)/(y, y)$. Evaluating $(x + ty, x + ty)$ at this $t = -(x, y)/(y, y)$ gives

$$-\frac{(x, y)^2}{(y, y)} + (x, x), \text{ which is } \geq 0.$$

This proves the Cauchy-Schwarz inequality when $y \neq 0$. But the case of $y = 0$ is trivial.

The triangle inequality follows from the Cauchy-Schwarz inequality by

$$\|x+y\|^2 = (x+y, x+y) = (x, x)+2(x, y)+(y, y) \leq (x, x)+2\|x\|\|y\|+(y, y) = (\|x\| + \|y\|)^2.$$

The Pythagorean relation clearly follows from the above line of proof when $(x, y) = 0$. ■

In the context of $\mathcal{C}[a, b]$, the Cauchy-Schwarz inequality takes the form of

$$|\int_a^b f(x)g(x)\,dx| \leq \sqrt{\int_a^b |f(x)|^2\,dx}\sqrt{\int_a^b |g(x)|^2\,dx}$$

for $f, g \in \mathcal{C}[a, b]$, and the triangle inequality takes the form of

$$\sqrt{\int_a^b |f(x) + g(x)|^2\,dx} \leq \sqrt{\int_a^b |f(x)|^2\,dx} + \sqrt{\int_a^b |g(x)|^2\,dx}.$$

Exercise 4.1.4 A Set of Functions Orthogonal on $[0, l]$. Verify that the set of functions $\left\{\sin\left(\frac{n\pi x}{l}\right) : n \in \mathbb{N}\right\}$ are mutually orthogonal to each other on $[0, l]$ and that

$$\| \sin\left(\frac{n\pi x}{l}\right) \| = \sqrt{\frac{l}{2}} \text{ for } n \in \mathbb{N}.$$

Then show that

$$\int_0^l |\sum_{n=1}^N b_n \sin\left(\frac{n\pi x}{l}\right)|^2\, dx = \frac{l}{2}\left(\sum_{n=1}^N |b_n|^2\right).$$

Hint. Use the relation $2\sin A \sin B = \cos(A-B) - \cos(A+B)$.

Exercise 4.1.5 A Set of Functions Orthogonal on $[-l,l]$. Verify that the set of functions $\left\{1, \cos\left(\frac{n\pi x}{l}\right), \sin\left(\frac{n\pi x}{l}\right) : n \in \mathbb{N}\right\}$ are mutually orthogonal to each other on $[-l,l]$ and that

$$\|1\| = \sqrt{2l}, \| \cos\left(\frac{n\pi x}{l}\right)\| = \| \sin\left(\frac{n\pi x}{l}\right)\| = \sqrt{l} \text{ for } n \in \mathbb{N}.$$

Then show that

$$\int_0^{2l} |a_0 + \sum_{n=1}^N \left(a_n \cos\left(\frac{n\pi x}{l}\right) + b_n \sin\left(\frac{n\pi x}{l}\right)\right)|^2\, dx = l\left(2|a_0|^2 + \sum_{n=1}^N \left(|a_n|^2 + |b_n|^2\right)\right).$$

Hint. Use the relations $2\sin A \sin B = \cos(A-B) - \cos(A+B), 2\cos A \cos B = \cos(A-B) + \cos(A+B), 2\sin A \cos B = \sin(A+B) - \sin(A-B)$.

It is often necessary and productive to work with spaces of complex valued functions, which should be regarded as vector spaces over $\mathbb{C}$. The notion of an inner product can be extended to a vector space over $\mathbb{C}$, with some modification.

Definition 4.1.6 Hermitian Inner Product Space.

A vector space V over $\mathbb{C}$ is called a Hermitian inner product space if there is a function $(x,y) \in V \times V \mapsto (x,y) \in \mathbb{C}$ such that

1. $(a_1x_1 + a_2x_2, y) = a_1(x_1,y) + a_2(x_2,y)$ for any $x_1, x_2, y \in V$ and any $a_1, a_2 \in \mathbb{C}$;
2. $(x,y) = \overline{(y,x)}$ for any $x, y \in V$;
3. (x,x) is a nonnegative real number for any $x \in V$ and equals 0 iff $x = 0$.

Two vectors $x, y \in V$ are (Hermitian) orthogonal if $(x,y) = 0$.

Note that the first two properties imply

$$(x, a_1y_1 + a_2y_2) = \overline{(a_1y_1 + a_2y_2, x)} = \overline{a_1}(x,y_1) + \overline{a_2}(x,y_2) \text{ for any } x, y_1, y_2 \in V.$$

Note that a Hermitian inner product on a vector space is not bilinear in both variables; it is linear in the first variable, but complex conjugate linear in the second variable.

The Cauchy-Schwarz and triangle inequalities and the notion of norm induced by the inner product extend readily to a Hermitian inner product space.

To distinguish between a Hermitian inner product and an inner product introduced earlier on a vector space over the reals, we will refer to the latter as a Euclidean inner product.

For complex valued functions f, g in $\mathcal{C}[a,b]$, a natural Hermitian inner product is $(f,g) = \int_a^b f(x)\overline{g(x)}\, dx$. This is consistent with the inner product introduced earlier on $\mathcal{C}[a,b]$ when f, g are real valued.

Exercise 4.1.7 The Orthogonal Family $\{e^{i\frac{n\pi x}{l}} : n \in \mathbb{Z}\}$ on $[-l, l]$. Verify that the set of functions $\{e^{i\frac{n\pi x}{l}} : n \in \mathbb{Z}\}$ are orthogonal on $[-l, l]$ and that

$$\|e^{i\frac{n\pi x}{l}}\| = \sqrt{2l} \text{ for all } n \in \mathbb{Z}.$$

Then show that

$$\int_{-l}^{l} |\sum_{-N}^{N} c_n e^{i\frac{n\pi x}{l}}|^2 \, dx = 2l \left(\sum_{-N}^{N} |c_n|^2 \right).$$

Remark 4.1.8

Two modifications are made in the defining properties of a Hermitian inner product: (i). allowing (x, y) to take complex values, and (ii). replacing the symmetry property by the complex conjugate symmetry property $(x, y) = \overline{(y, x)}$.

These are based on the following considerations. (a). It is preferable to keep some complex linearity for an inner product on a vector space over $\mathbb{C}$ such as given in the first property of an inner product, and this makes it necessary to allow (x, y) to take complex values. (b). We still would like to use $\sqrt{(x, x)}$ as a norm for a vector, thus we need (x, x) to be a nonnegative real number for any $x \in V$.

Let $H(x, y), S(x, y)$ denote, respectively, the real and imaginary parts of (x, y):

$$(x, y) = G(x, y) + iS(x, y), x, y \in V.$$

Then we need $S(x, x) = 0$ for all $x \in V$. This property, coupled with linearity over $\mathbb{R}$, implies that $S(x, y)$ must be antisymmetric in x, y. In fact, it makes sense to require $G(x, y)$ to be an inner product on V treating V as a vector space over $\mathbb{R}$. Thus we want $G(x, y)$ to be symmetric in x, y. An additional desired property is that multiplication by i on both x and y should preserve the inner product:

$$(ix, iy) = (x, y) \text{ for all } x, y \in V.$$

It turns out that these desired properties are encoded in, in fact, equivalent to the defining properties for a Hermitian inner product.

Exercise 4.1.9 Hermitian and Euclidean Inner Product. Use the set up of $G(x, y) + iS(x, y) = (x, y)$ for a Hermitian inner product. Verify that

1. $G(x, y) = G(y, x), S(x, y) = -S(y, x)$ for all $x, y \in V$.

2. $G(ix, iy) = G(x, y), S(ix, iy) = S(x, y)$ for all $x, y \in V$.

3. $S(x, y) = G(x, iy)$ for all $x, y \in V$.

4. $G(x, ix) = 0$ for all $x \in V$.

Conversely, if $G(x, y)$ is an inner product on V as a vector space over the reals, and satisfies $G(ix, iy) = G(x, y)$ for all $x, y \in V$. Then define $S(x, y) = G(x, iy)$ and $(x, y) = G(x, y) + iS(x, y)$ for all $x, y \in V$. Verify that this (x, y) is a Hermitian inner product on V. In other words, a Hermitian inner product is always associated with a Euclidean inner product which is preserved by multiplication by i.

Remark 4.1.10

When two vectors x, y are orthogonal in a complex vector space V with a Hermitian inner product, the Pythagorean relation holds in the same way as for case in a vector space with a Euclidean inner product. In this sense it is a proper generalization of the orthogonal relation in Euclidean geometry. On the other hand, there is a subtle difference: suppose $\{\mathbf{v}_1, \cdots, \mathbf{v}_m\}$ is an orthonormal basis of a complex vector space V with a Hermitian inner product, then $\mathbf{v}_j$ and $i\mathbf{v}_j$ are not orthogonal in this Hermitian inner product; on the other hand, if we endow V with the indued Euclidean inner product $G(x, y)$ as discussed above, then $\mathbf{v}_j$ and $i\mathbf{v}_j$ are orthogonal in $G(x, y)$; in fact, $\{\mathbf{v}_1, \cdots, \mathbf{v}_m; i\mathbf{v}_1, \cdots, i\mathbf{v}_m\}$ becomes an orthonormal basis for this inner product $G(x, y)$. This is the case for the canonical Hermitian metric on $\mathbb{C}$, where $(z, w) = z\bar{w}$, so $(1, i) = -i \neq 0$, while in the geometric representation of complex numbers, $1, i$ are orthogonal---they are orthogonal in the induced G inner product, which is the standard Euclidean inner product.

In the following we will not distinguish between a Hermitian inner product and a Euclidean inner product, and will let the context to imply the appropriate one.

Definition 4.1.11 Orthonormal Vectors.

A set of vectors $\{\mathbf{v}_1, \mathbf{v}_2, \cdots\}$ (finite or infinite) in an inner product space V is called an orthonormal set, if any two distinct vectors in this set are orthogonal to each other and each one is a unit vector.

Definition 4.1.12 Fourier Coefficients and Fourier Series.

Let $\{\mathbf{v}_1, \mathbf{v}_2, \cdots\}$ be a set of orthonormal vectors in an inner product space V. For any vector $\mathbf{v} \in V$, define $c_k = (\mathbf{v}, \mathbf{v}_k)$. Then $\{c_k\}$ are called the Fourier coefficients of $\mathbf{v}$ with respect to this set of orthonormal vectors, and the series $\sum_k c_k \mathbf{v}_k$ is called the Fourier series of $\mathbf{v}$ with respect to this set of orthonormal vectors.

In the above definition the convergence of $\sum_k c_k \mathbf{v}_k$ in the case of an infinite set of orthonormal vectors is not directly addressed; one either needs to show that the series converges or simply assumes it as a formal series at this point. As will be seen soon, it is also appropriate to call this sum the orthogonal projection of $\mathbf{v}$ in the span of this set of orthonormal vectors. Note that in a setting of a set of infinite vectors, we take the span of such a set to mean the **completion** of the space of *finite* linear combination of vectors from this set, which allows us to make sense of an infinite series of such vectors.

Remark 4.1.13

Sometimes we work with a set of orthogonal vectors which are not necessarily unit vectors. This is the case with $\left\{1, \cos\left(\frac{n\pi x}{l}\right), \sin\left(\frac{n\pi x}{l}\right) : n \in \mathbb{N}\right\}$ on $[-l, l]$. Here we modify the definition of the Fourier coefficients of $g \in \mathcal{R}[-l, l]$ by defining

$$a_0 = \frac{(g, 1)}{(1, 1)} = \frac{1}{2l} \int_{-l}^{l} g(x)\, dx,$$

$$a_n = \frac{\left(g, \cos\left(\frac{n\pi x}{l}\right)\right)}{\left(\cos\left(\frac{n\pi x}{l}\right), \cos\left(\frac{n\pi x}{l}\right)\right)} = \frac{1}{l}\int_{-l}^{l} g(x)\cos\left(\frac{n\pi x}{l}\right)\,dx \text{ for } n \geq 1,$$

$$b_n = \frac{\left(g, \sin\left(\frac{n\pi x}{l}\right)\right)}{\left(\sin\left(\frac{n\pi x}{l}\right), \sin\left(\frac{n\pi x}{l}\right)\right)} = \frac{1}{l}\int_{-l}^{l} g(x)\sin\left(\frac{n\pi x}{l}\right)\,dx \text{ for } n \geq 1,$$

and

$$S_N[g](x) := a_0 + \sum_{n=1}^{N}\left[a_n \cos\left(\frac{n\pi x}{l}\right) + b_n \sin\left(\frac{n\pi x}{l}\right)\right]$$

as the partial sums of the Fourier series of $g(x)$ on $[-l, l]$:

$$g(x) \sim a_0 + \sum_{n=1}^{\infty}\left[a_n \cos\left(\frac{n\pi x}{l}\right) + b_n \sin\left(\frac{n\pi x}{l}\right)\right].$$

When we work with the set of orthogonal functions $\left\{\sin\left(\frac{n\pi x}{l}\right) : n \in \mathbb{N}\right\}$ on $[0, l]$, the Fourier coefficients of $g \in \mathcal{R}[0, l]$ with respect to this set of orthogonal functions on $[0, l]$ are defined by

$$b_n = \frac{\left(g, \sin\left(\frac{n\pi x}{l}\right)\right)}{\left(\sin\left(\frac{n\pi x}{l}\right), \sin\left(\frac{n\pi x}{l}\right)\right)} = \frac{2}{l}\int_{0}^{l} g(x)\sin\left(\frac{n\pi x}{l}\right)\,dx.$$

Exercise 4.1.14 Find Fourier Coefficients. Find the Fourier coefficients of the functions $f(x) = 1$ and $g(x) = \cos x$ with respect to the set of orthogonal functions $\{\sin(nx) : n \in \mathbb{N}\}$ on $[0, \pi]$.

Exercise 4.1.15 Integral Representation of Fourier Partial Sums. Let

$$S_N[g](x) := a_0 + \sum_{n=1}^{N}\left[a_n \cos\left(\frac{n\pi x}{l}\right) + b_n \sin\left(\frac{n\pi x}{l}\right)\right]$$

denote the partial sums of the Fourier series of $g(x)$ with respect to the set of orthogonal functions $\left\{1, \cos\left(\frac{n\pi x}{l}\right), \sin\left(\frac{n\pi x}{l}\right) : n \in \mathbb{N}\right\}$ on $[-l, l]$, and let

$$c_n = \frac{1}{2l}\int_{-l}^{l} g(x)e^{-\frac{in\pi x}{l}}\,dx$$

denote the Fourier coefficients of $g(x)$ with respect to the set of orthogonal functions $\left\{e^{\frac{in\pi x}{l}}\right\}_{n=-N}^{N}$. Verify that

$$S_N[g](x) = \sum_{n=-N}^{N} c_n e^{\frac{in\pi x}{l}} = \frac{1}{2l}\int_{-l}^{l} g(t)D_N(x-t)\,dt$$

where

$$D_N(t) = \sum_{-N}^{N} e^{-\frac{in\pi t}{l}} = \frac{\sin\frac{(N+\frac{1}{2})\pi t}{l}}{\sin\frac{\pi t}{2l}}.$$

Hint. First need to work out

$$\begin{aligned} S_N[g](x) &= \frac{1}{2l}\int_{-l}^{l} g(t)\left(1 + \sum_{n=1}^{N} 2\cos\left(\frac{n\pi(x-t)}{l}\right)\right)dt \\ &= \frac{1}{2l}\int_{-l}^{l} g(t)\left(\sum_{n=-N}^{N} e^{\frac{in\pi(x-t)}{l}}\right)dt, \end{aligned}$$

then establish

$$1+\sum_{n=1}^{N} 2\cos\left(\frac{n\pi s}{l}\right) = \sum_{n=-N}^{N} e^{\frac{in\pi s}{l}} = \frac{\sin\frac{(N+\frac{1}{2})\pi s}{l}}{\sin\frac{\pi s}{2l}}$$

using either the relation $2\sin\frac{\pi s}{2l}\cos\frac{(N+\frac{1}{2})\pi s}{l} = \sin(\frac{(N+1)\pi s}{l}) - \sin(\frac{N\pi s}{l})$ or $\cos\left(\frac{n\pi s}{l}\right) = \mathrm{Re}(e^{\frac{in\pi s}{l}})$.

Theorem 4.1.16 Best Approximation Property of Fourier Series.

Let $\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_N\}$ be a finite set of orthonormal vectors in an inner product space V. For any vector $\mathbf{v} \in V$, let $\sum_{k=1}^{N} c_k\mathbf{v}_k$ be the Fourier series of $\mathbf{v}$ with respect to this set of orthonormal vectors. Then

$$(\mathbf{v} - \sum_{k=1}^{N} c_k\mathbf{v}_k, \mathbf{v}_j) = 0 \textit{ for all } j = 1, \cdots, N$$

and $\mathbf{v} - \sum_{k=1}^{N} c_k\mathbf{v}_k$ is orthogonal to every vector in the span of this set of orthonormal vectors. Furthermore,

$$\|\mathbf{v}\|^2 = \|\mathbf{v} - \sum_{k=1}^{N} c_k\mathbf{v}_k\|^2 + \|\sum_{k=1}^{N} c_k\mathbf{v}_k\|^2,$$

and

$$\|\mathbf{v} - \sum_{k=1}^{N} c_k\mathbf{v}_k\| \leq \|\mathbf{v} - \sum_{k=1}^{N} a_k\mathbf{v}_k\| \textit{ for any coefficients } \{a_k\},$$

namely, $\sum_{k=1}^{N} c_k\mathbf{v}_k$ is closest to $\mathbf{v}$ among all vectors in the span of this set of orthonormal vectors.

Proof. The first assertion follows directly from

$$(\mathbf{v} - \sum_{k=1}^{N} c_k\mathbf{v}_k, \mathbf{v}_j) = (\mathbf{v}, \mathbf{v}_j) - \sum_{k=1}^{N} c_k(\mathbf{v}_k, \mathbf{v}_j) = c_j - c_j = 0$$

using the orthonormal condition $(\mathbf{v}_k, \mathbf{v}_j) = \delta_{kj}$.

The second assertion follows by using the orthogonality relations above

$$\begin{aligned}\|\mathbf{v}\|^2 &= \left((\mathbf{v} - \sum_{k=1}^{N} c_k\mathbf{v}_k) + \sum_{k=1}^{N} c_k\mathbf{v}_k, (\mathbf{v} - \sum_{k=1}^{N} c_k\mathbf{v}_k) + \sum_{k=1}^{N} c_k\mathbf{v}_k\right)\\ &= (\mathbf{v} - \sum_{k=1}^{N} c_k\mathbf{v}_k, \mathbf{v} - \sum_{k=1}^{N} c_k\mathbf{v}_k) + 2(\mathbf{v} - \sum_{k=1}^{N} c_k\mathbf{v}_k, \sum_{k=1}^{N} c_k\mathbf{v}_k) + (\sum_{k=1}^{N} c_k\mathbf{v}_k, \sum_{k=1}^{N} c_k\mathbf{v}_k)\\ &= \|\mathbf{v} - \sum_{k=1}^{N} c_k\mathbf{v}_k\|^2 + \|\sum_{k=1}^{N} c_k\mathbf{v}_k\|^2\end{aligned}$$

Set $\mathbf{w} = \sum_{k=1}^{N} c_k\mathbf{v}_k$. Then $(\mathbf{v} - \mathbf{w}, \sum_{k=1}^{N}(c_k - a_k)\mathbf{v}_k) = 0$, so

$$\|\mathbf{v} - \sum_{k=1}^{N} a_k\mathbf{v}_k\|^2 = \left(\mathbf{v} - \mathbf{w} + (\sum_{k=1}^{N}(c_k - a_k)\mathbf{v}_k), \mathbf{v} - \mathbf{w} + (\sum_{k=1}^{N}(c_k - a_k)\mathbf{v}_k)\right)$$

$$\begin{aligned}
&= (\mathbf{v}-\mathbf{w}, \mathbf{v}-\mathbf{w}) + \left(\sum_{k=1}^{N} (c_k - a_k)\mathbf{v}_k, \sum_{k=1}^{N} (c_k - a_k)\mathbf{v}_k \right) \\
&= \|\mathbf{v}-\mathbf{w}\|^2 + \| \sum_{k=1}^{N} (c_k - a_k)\mathbf{v}_k \|^2 \\
&\geq \|\mathbf{v}-\mathbf{w}\|^2,
\end{aligned}$$

with equality iff $c_k - a_k = 0$ for all $k, 1 \leq k \leq N$. ■

Definition 4.1.17 Orthogonal Projection.

Let $\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_N\}$ be a finite set of orthonormal vectors in an inner product space V. For any vector $\mathbf{v} \in V$, let $\sum_{k=1}^{N} c_k \mathbf{v}_k$ be the Fourier series of $\mathbf{v}$ with respect to this set of orthonormal vectors. Then $\sum_{k=1}^{N} c_k \mathbf{v}_k$ is also called the orthogonal projection of $\mathbf{v}$ in the span of this set of orthonormal vectors.

Exercise 4.1.18 Find Orthogonal Projections. Find the orthogonal projections of the functions $f(x) = 1$ and $g(x) = \cos x$ in the span of the set of orthogonal functions $\{\sin(nx) : 1 \leq n \leq N\}$ on $[0, \pi]$.

Exercise 4.1.19 Find the orthogonal projection of the function $f(x) = x$ in the span of the set of orthogonal functions $\{1, \cos(nx), \sin(nx) : 1 \leq n \leq N\}$ on $[-\pi, \pi]$.

Theorem 4.1.20 Bessel's Inequality.

Let $\{\mathbf{v}_1, \mathbf{v}_2, \cdots\}$ be a set of orthonormal vectors in an inner product space V. For any vector $\mathbf{v} \in V$, let $\sum_k c_k \mathbf{v}_k$ be the Fourier series of $\mathbf{v}$ with respect to this set of orthonormal vectors. Then

$$\sum_k |c_k|^2 \leq \|\mathbf{v}\|^2 \qquad \textit{(Bessel's inequality)}.$$

Proof. Take any finite subset $\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_N\}$. Then we have already proved that

$$\|\mathbf{v}\|^2 \geq \| \sum_{k=1}^{N} c_k \mathbf{v}_k \|^2 = \sum_{k=1}^{N} |c_k|^2.$$

Since this holds for any finite N, Bessel's inequality follows immediately. ■

Corollary 4.1.21 Bessel's inequality for trigonometric Fourier series.

Let

$$g(x) \sim a_0 + \sum_{n=1}^{\infty} \left[a_n \cos\left(\frac{n\pi x}{l}\right) + b_n \sin\left(\frac{n\pi x}{l}\right) \right].$$

be the Fourier series of $g \in \mathcal{R}[-l, l]$ with respect to the set of orthogonal functions $\left\{1, \cos\left(\frac{n\pi x}{l}\right), \sin\left(\frac{n\pi x}{l}\right) : n \in \mathbb{N}\right\}$ on $[-l, l]$. Then

$$\int_{-l}^{l} |g(x)|^2 \, dx = \int_{-l}^{l} |g(x) - S_N[g](x)|^2 \, dx + \int_{-l}^{l} |S_N[g](x)|^2 \, dx \qquad (4.1.1)$$

$$= \int_{-l}^{l} |g(x) - S_N[g](x)|^2\, dx + 2l|a_0|^2 + l \sum_{n=1}^{N} \left[|a_n|^2 + |b_n|^2\right]. \tag{4.1.2}$$

As a consequence the following Bessel's inequality holds

$$\int_{-l}^{l} |g(x)|^2\, dx \geq 2l|a_0|^2 + l \sum_{n=1}^{\infty} \left[|a_n|^2 + |b_n|^2\right].$$

Remark 4.1.22

So far we have not addressed the issue whether $S_N[g](x)$ converges to $g(x)$ in the mean square sense, namely whether $\int_{-l}^{l} |g(x) - S_N[g](x)|^2\, dx \to 0$ as $N \to \infty$. From the above we see that the answer depends on whether equality holds in the Bessel's inequality. Another possible approach is to show that there exists a sequence $p_N(x) = \sum_{n=-N}^{N} c'_n e^{\frac{inx}{l}}$ such that $\int_{-l}^{l} |g(x) - p_N(x)|^2\, dx \to 0$ as $N \to \infty$ and appeal to the Best Approximation Property of the Fourier series. This would be the case if the set of finite linear combination of functions from $\{e^{\frac{inx}{l}}\}$ is dense in the mean square norm in the set of function spaces, such as $\mathcal{R}[-l,l]$ or $C[-l,l]$, in which we are interested in making such a Fourier series expansion.

Given any g in $\mathcal{R}[-l,l]$ or $C[-l,l]$, using the orthogonality relations we have

$$\int_{-l}^{l} |S_{N'}[g](x) - S_N[g](x)|^2\, dx = l \sum_{n=N}^{N'} \left[|a_n|^2 + |b_n|^2\right]$$

for $N' > N$. The Bessel's inequality implies then that the sequence $\{S_N[g](x)\}$ is a Cauchy sequence in the mean square norm. At this point we need the property of completeness of the function space on which we are working. Unfortunately, neither $\mathcal{R}[-l,l]$ nor $C[-l,l]$ is complete with respect to the mean square norm. The completion of either $\mathcal{R}[-l,l]$ or $C[-l,l]$ turns out to the space of Lebesgue square integrable functions $L^2[-l,l]$. So a more proper discussion on the issue of mean square convergence should be on the complete space $L^2[-l,l]$.

Without discussing details of Lebesgue integrable functions, we may assume that there exists some $\hat{g} \in L^2[-l,l]$ such that $\int_{-l}^{l} |\hat{g} - S_N[g](x)|^2\, dx \to 0$ as $N \to \infty$. We claim the following property:

$$(g - \hat{g}, e^{\frac{im\pi x}{l}}) = 0 \quad \text{for any } m \in \mathbb{Z}.$$

This follows by noting that for any $N > m$,

$$(g - S_N[g](x), e^{\frac{im\pi x}{l}}) = (g, e^{\frac{im\pi x}{l}}) - (S_N[g](x), e^{\frac{im\pi x}{l}}) = 2l\,(c_m - c_m) = 0,$$

so

$$(g-\hat{g}, e^{\frac{im\pi x}{l}}) = (g-S_N[g](x), e^{\frac{im\pi x}{l}}) + (S_N[g](x)-\hat{g}, e^{\frac{im\pi x}{l}}) = (S_N[g](x)-\hat{g}, e^{\frac{im\pi x}{l}}),$$

while by the Cauchy-Schwarz inequality

$$|(S_N[g](x) - \hat{g}, e^{\frac{im\pi x}{l}})| \leq \|S_N[g](x) - \hat{g}\| \|e^{\frac{im\pi x}{l}}\| \to 0$$

as $N \to \infty$, which shows that

$$(g - \hat{g}, e^{\frac{im\pi x}{l}}) = 0.$$

So a key to our question is whether the only function which is orthogonal to all $e^{\frac{im\pi x}{l}}$ is the 0 function. This discussion makes sense on a general inner product space.

Definition 4.1.23 A Complete (or Maximal) Orthonormal Set.

A set of orthonormal vectors in an inner product space is called complete (or maximal) if the only vector orthogonal to each of these vectors is the zero vector.

4.2 Convergence of the Trigonometric Fourier Series

We now investigate the issue of convergence of the trigonometric Fourier series. To simplify the set up, we will take $l = \pi$ from now on; the general case can be reduced to this case by a simple change of variables $x \mapsto \pi x/l$.

In an earlier exercise it is established that

$$S_N[g](x) = \sum_{n=-N}^{N} c_n e^{inx} = \frac{1}{2\pi}\int_{-\pi}^{\pi} g(t)D_N(x-t)\,dt$$

where

$$D_N(t) = \sum_{-N}^{N} e^{-int} = \frac{\sin(N+\frac{1}{2})t}{\sin\frac{t}{2}}.$$

Note that

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} D_N(t)\,dt = 1.$$

Using this property and extending $g(x)$ as a 2π periodic function on $\mathbb{R}$ we have

$$S_N[g](x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} g(x-t)D_N(t)\,dt$$

by a change of variable $t \mapsto x - t$ in $\int_{-\pi}^{\pi} g(t)D_N(x-t)\,dt$. Then

$$\begin{aligned} S_N[g](x) - g(x) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} [g(x-t) - g(x)]D_N(t)\,dt \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} h(x;t)\sin(N+\frac{1}{2})t\,dt, \end{aligned}$$

where $h(x;t) = \frac{g(x-t)-g(x)}{\sin\frac{t}{2}}$. The following Riemann-Lebesgue Lemma plays an important role.

Lemma 4.2.1 Riemann-Lebesgue Lemma.

For any Riemann integrable function h on $(-\pi,\pi)$, the following holds

$$\int_{-\pi}^{\pi} h(t)\sin(\lambda t)\,dt \to 0 \text{ as } \lambda \to \infty.$$

Proof. If we only take $\lambda = N$ as integers, then this follows from the Bessel's inequality. The same argument also shows that $\int_{-\pi}^{\pi} h(t)\cos(Nt)\,dt \to 0$ as $N \to \infty$. These properties are sufficient for applications to $\int_{-\pi}^{\pi} h(x;t)\sin(N+\frac{1}{2})t\ dt$, as $\sin(N+\frac{1}{2})t = \cos\frac{t}{2}\sin(Nt) + \sin\frac{t}{2}\cos(Nt)$.

For the general case, first note that if $h \in C^1[-\pi,\pi]$, then integration by parts gives

$$\int_{-\pi}^{\pi} h(t)\sin(\lambda t)\,dt = -h(t)\lambda^{-1}\cos(\lambda t)\Big|_{-\pi}^{\pi} + \int_{-\pi}^{\pi} \lambda^{-1}\cos(\lambda t)h'(t)\,dt \to 0$$

as $\lambda \to \infty$. For the given $h \in \mathcal{R}[-\pi,\pi]$, take any $\epsilon > 0$, we first find some $\hat{h} \in C^1[-\pi,\pi]$ such that $\int_{-\pi}^{\pi} |h(t) - \hat{h}(t)|\,dt < \epsilon$. Then

$$\int_{-\pi}^{\pi} h(t)\sin(\lambda t)\,dt = \int_{-\pi}^{\pi} [h(t) - \hat{h}(t)]\sin(\lambda t)\,dt + \int_{-\pi}^{\pi} \hat{h}(t)\sin(\lambda t)\,dt,$$

$$|\int_{-\pi}^{\pi} [h(t) - \hat{h}(t)]\sin(\lambda t)\,dt| \le \int_{-\pi}^{\pi} |h(t) - \hat{h}(t)|\,dt < \epsilon,$$

and $|\int_{-\pi}^{\pi} \hat{h}(t)\sin(\lambda t)\,dt| < \epsilon$ for all sufficiently large λ, which concludes our proof. ■

Theorem 4.2.2 A Convergence Criterion for the Trigonometric Fourier series.

Suppose that g is a 2π periodic function on $\mathbb{R}$ and $x \in (-\pi,\pi)$ is such that there exist $\delta > 0$ and $M < \infty$ such that

$$|g(x+t) - g(x)| \le M|t| \text{ for all } t \in (-\delta,\delta). \tag{4.2.1}$$

Then $S_N[g](x) - g(x) \to 0$ as $N \to \infty$.

Proof. Under our assumption, $h(x;t) = \frac{g(x-t)-g(x)}{\sin\frac{t}{2}}$ is Riemann integrable on $(-\pi,\pi)$, so we can apply the Riemann-Lebesgue Lemma 4.2.1 to draw our conclusion. ■

Remark 4.2.3

Condition (4.2.1) implies that $g(t)$ is continuous at x. Note that if $g'(x)$ exists, then (4.2.1) is satisfied at x. In fact, (4.2.1) is satisfied at x if both the left derivative $g'(x-)$ and the right derivative $g'(x+)$ exist.

If $g(t)$ has both a right limit $g(x+)$ and a left limit $g(x-)$ at x and $g(x+) \ne g(x-)$, we can rewrite $S_N[g](x)$ as

$$S_N[g](x) = \frac{1}{2\pi}\int_0^{\pi} [g(x-t) + g(x+t)]\,D_N(t)\,dt,$$

so

$$\begin{aligned} &S_N[g](x) - \frac{g(x+)+g(x-)}{2} \\ =&\frac{1}{2\pi}\int_0^{\pi} [g(x-t) - g(x-) + g(x+t) - g(x+)]\,D_N(t)\,dt. \end{aligned}$$

If $g(t)$ satisfies

$$|g(x-t) - g(x-)|, |g(x+t) - g(x+)| \le Mt \text{ for } t > 0 \text{ near } 0,$$

we can make the same argument to show that $S_N[g](x) - \frac{g(x+)+g(x-)}{2} \to 0$ as $N \to \infty$.

Note that whether or not $h(x;t)$ is integrable so as to apply Lemma 4.2.1 depends only on the the local behavior of $g(t)$ near x, so whether or not $S_N[g](x)$ converges depends only on the local behavior of $g(t)$ near x. This is known as the Riemann's localization theorem.

As a consequence of the above theorem, if $g \in C[-\pi, \pi]$ and has continuous derivative on $C[-\pi, \pi]$, we can redefine $g(-\pi)$ to be $g(\pi)$ and extend this function to be a 2π periodic function on $\mathbb{R}$. Then the resulting function satisfies the assumption of the above theorem at any x such that $-\pi < x < \pi$, so $S_N[g](x) \to g(x)$ as $N \to \infty$. At $x = \pi$, the left limit of the extended function is $g(\pi)$ and the right limit of the extended function is $g(-\pi)$. Thus $S_N[g](\pi) \to \frac{g(\pi)+g(-\pi)}{2}$ as $N \to \infty$. Since the Fourier series expansion here is 2π periodic, the proper interpretation of the expansion is that the series equals the 2π periodic extension of the given function on $(-\pi, \pi]$ in the above sense.

Remark 4.2.4

For any $x \in (-\pi, \pi)$, there exist continuous functions f such that $S_N[f](x)$ does not converge, so continuity of f at x alone may not guarantee that $S_N[f](x)$ converges as $N \to \infty$.

However, a modified trigonometric series, the ***Fejér*** *series, defined as the arithmetic average of $S_n[f](x)$'s:*

$$\sigma_N[f](x) = \frac{S_0[f](x) + S_1[f](x) + \cdots + S_N[f](x)}{N+1}$$

does converge to $f(x)$ for any f which is continuous at x.

Theorem 4.2.5 Fejér Theorem.

Suppose that the Riemann integrable function f on $(-\pi, \pi)$ is continuous at $x \in (-\pi, \pi)$, then $\sigma_N[f](x) \to f(x)$ as $N \to \infty$. If f is continuous on $[-\pi, \pi]$ and is 2π periodic, then $\sigma_N[f](x) \to f(x)$ uniformly over $[-\pi, \pi]$ as $N \to \infty$. Suppose that $f(t)$ has both a right limit $f(x+)$ and a left limit $f(x-)$ at x, then $\sigma_N[f](x) \to \frac{f(x+)+f(x-)}{2}$ as $N \to \infty$.

Proof. We note that

$$\sigma_N[f](x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x-t) K_N(t)\, dt$$

where

(a)

$$K_N(t) := \frac{D_0(t) + D_1(t) + \cdots + D_N(t)}{N+1} = \frac{1 - cos(N+1)t}{(N+1)(1-\cos t)} \geq 0,$$

(b)

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} K_N(t)\, dt = 1 \text{ for any } N,$$

(c)

$$\text{for any } 0<\delta<\pi, \int_{\delta\le|x|\le\pi} K_N(t)\,dt \to 0 \text{ as } N\to\infty.$$

For, then, using (a) above,

$$\begin{aligned}
&|\sigma_N[f](x)-f(x)|\\
=&|\frac{1}{2\pi}\int_{-\pi}^{\pi}[f(x-t)-f(x)]K_N(t)\,dt|\\
\le&\frac{1}{2\pi}\int_{-\delta}^{\delta}|f(x-t)-f(x)|K_N(t)\,dt+\frac{\max|f|}{\pi}\int_{\delta\le|t|\le\pi}K_N(t)\,dt,
\end{aligned}$$

and for any $\epsilon>0$, we first choose $0<\delta<\pi$ such that $|f(x-t)-f(x)|<\epsilon$ for all t with $|t|\le\delta$, then use this δ in the above, which will make the first integral above $<\epsilon$ using (b) above. Finally, for sufficiently large N, the second integral above will also be less than ϵ using (c) above. This proves the first assertion. When f is continuous on $[-\pi,\pi]$ and is 2π periodic, then δ above can be chosen independent of $x\in[-\pi,\pi]$, which shows that the convergence is uniform over $[-\pi,\pi]$.

If f has both left and right limits at x, one could use the evenness of $K_N(t)$ or $D_N(t)$ to rewrite the integral $\int_{-\pi}^{\pi}$ as two separate integrals. For example,

$$\begin{aligned}
\sigma_N[f](x)&=\frac{1}{2\pi}\int_{-\pi}^{\pi}f(x-t)K_N(t)\,dt\\
&=\frac{1}{2\pi}\int_{0}^{\pi}[f(x-t)+f(x+t)]\,K_N(t)\,dt
\end{aligned}$$

and one can use

$$\frac{1}{2\pi}\int_0^{\pi}K_N(t)\,dt=\frac{1}{2}$$

to carry out a similar convergence argument. ■

Fejér's theorem implies that the span of $\{e^{inx}\}$ (equivalently $\{\cos(nx),\sin(nx)\}$) is dense in both $C[-\pi,\pi]$ and $\mathcal{R}[-\pi,\pi]$ in the mean square sense, for, given any $g\in\mathcal{R}[-\pi,\pi]$ and $\epsilon>0$, first find a 2π periodic and continuous function f such that $\|f-g\|:=\left(\int_{-\pi}^{\pi}|f(t)-g(t)|^2\,dt\right)^{1/2}<\epsilon$. Then find N such that $\|f-\sigma_N(f)\|<\epsilon$ by the above theorem. Finally triangle inequality implies that $\|g-\sigma_N(f)\|<2\epsilon$.

As a consequence we have

Theorem 4.2.6 Mean Square Convergence and Parseval Equality.

For any $g\in\mathcal{R}[-\pi,\pi]$, $\|S_N(g)-g\|\to 0$ as $N\to\infty$. As a consequence, the following Parseval equality holds:

$$\int_0^{2l}|g(x)|^2\,dx=\pi\left(2|a_0|^2+\sum_{n=1}^{\infty}\left[|a_n|^2+|b_n|^2\right]\right).$$

Proof. Since the span of $\{e^{inx}\}_{n=-N'}^{N'}$ is a subspace of $\{e^{inx}\}_{n=-N}^{N}$ when $N'\le N$, the Best Approximation Theorem implies that

$$\|S_N(g)-g\|\le\|S_{N'}(g)-g\| \text{ when } N'\le N.$$

For any $\epsilon>0$, Fejér's theorem gives some trigonometric polynomial p of degree N' such that $\|g-p\|<\epsilon$. Then the Best Approximation Theorem implies that

$\|S_{N'}(g) - g\| < \epsilon$. Then for all $N \geq N'$, we have $\|S_N(g) - g\| \leq \epsilon$. The Parseval equality follows from (4.1.2). ■

Theorem 4.2.7

$\{e^{inx}\}_{n=-\infty}^{\infty}$ is a set of maximal orthogonal functions in $L^2[-\pi, \pi]$

Proof. Suppose not. Let $g \in L^2[-\pi, \pi]$ be a non-zero function in $L^2[-\pi, \pi]$ orthogonal to each e^{inx}. We may assume that $\|g\| = 1$. Then for any trigonometric polynomial p of the form $\sum_{n=-N}^{N} c_n e^{inx}$, the orthogonality property implies that

$$\|g - p\|^2 = \|g\|^2 + \|p\|^2 \geq 1.$$

But we can find a continuous 2π periodic function f such that $\|g - f\| < \frac{1}{4}$, and a trigonometric polynomial p of the form $\sum_{n=-N}^{N} c_n e^{inx}$ such that $\|p - f\| < \frac{1}{4}$. Then the triangle inequality implies that

$$\|g - p\| \leq \|g - f\| + \|p - f\| < \frac{1}{2},$$

which contradicts the property $\|g - p\| \geq 1$ established earlier. ■

4.3 Term-wise Integration and Differentiation of Fourier Series

Suppose that

$$g(x) \sim a_0 + \sum_{n=1}^{\infty} \left[a_n \cos(nx) + b_n \sin(nx)\right].$$

is the Fourier series of $g \in \mathcal{R}[-\pi, \pi]$. We address whether we can integrate this relation term-wise and, when g' exists and is in $\mathcal{R}[-\pi, \pi]$, whether we can differentiate this relation term-wise.

Theorem 4.3.1 Term-wise Integration of Fourier Series.

Suppose that

$$g(x) \sim a_0 + \sum_{n=1}^{\infty} \left[a_n \cos(nx) + b_n \sin(nx)\right]$$

is the Fourier series of $g \in \mathcal{R}[-\pi, \pi]$. Then for any $a, b \in (-\pi, \pi)$,

$$\int_a^b g(x) = a_0(b - a) + \sum_{n=1}^{\infty} \int_a^b \left[a_n \cos(nx) + b_n \sin(nx)\right] dx.$$

Proof. The assertion is equivalent to

$$\int_a^b \left[g(x) - S_N(g; x)\right] dx \to 0 \text{ as } N \to \infty. \tag{4.3.1}$$

But

$$|\int_a^b \left[g(x) - S_N(g; x)\right] dx| \leq \int_a^b |g(x) - S_N(g; x)|\, dx$$

$$\leq (2\pi)^{1/2} \left(\int_{-\pi}^{\pi} |g(x) - S_N(g;x)|^2 \, dx \right)^{1/2},$$

and using $\int_{-\pi}^{\pi} |g(x) - S_N(g;x)|^2 \, dx \to 0$, we conclude (4.3.1). ■

Recall that a function is called piecewise continuous on $[a,b]$ if there is a finite partition $a = a_0 < a_1 < \cdots < a_m = b$ of $[a,b]$ such that its restriction on any (a_k, a_{k+1}) is continuous and has a continuous extension to $[a_k, a_{k+1}]$. This implies that the function is continuous at every point of $[a,b]$ with possibly the exception at the a_k's and that it has both left and right limits at each of these a_k's.

Theorem 4.3.2 Term-wise Differentiation of Fourier Series.

Suppose that $g(x)$ is continuous *on $[-\pi,\pi]$, and $g(-\pi) = g(\pi)$, and that $g'(x)$ exists except at a finite number of points and is piecewise continuous. Denote the Fourier series expansion of $g'(x)$ on $[-\pi,\pi]$ by*

$$g'(x) \sim a_0' + \sum_{n=1}^{\infty} \left[a_n' \cos(nx) + b_n' \sin(nx) \right].$$

Then

$$a_0' = 0, \text{ and } a_n' = nb_n, b_n' = -na_n.$$

In other words, under our assumptions here, *the Fourier series expansion of $g'(x)$ on $[-\pi,\pi]$ can be obtained by term-wise differentiation of the Fourier series expansion of $g(x)$ on $[-\pi,\pi]$.*

Proof. Note that, for $n=0$,

$$2\pi a_0' = \int_{-\pi}^{\pi} g'(x) \, dx = g(\pi) - g(-\pi) = 0,$$

and for $n\ge 1$, integration by parts gives

$$\begin{aligned}
\pi a_n' =& \int_{-\pi}^{\pi} g'(x) \cos(nx) \, dx \\
=& g(x) \cos(nx) \Big|_{x=-\pi}^{x=\pi} + n \int_{-\pi}^{\pi} g(x) \sin(nx) \, dx \text{ (using } g \in C[-\pi,\pi]) \\
=& n \int_{-\pi}^{\pi} g(x) \sin(nx) \, dx \text{ (using } g(-\pi) = g(\pi)) \\
=& \pi n b_n; \\
\pi b_n' =& \int_{-\pi}^{\pi} g'(x) \sin(nx) \, dx \\
=& g(x) \sin(nx) \Big|_{x=-\pi}^{x=\pi} - n \int_{-\pi}^{\pi} g(x) \cos(nx) \, dx \text{ (using } g \in C[-\pi,\pi]) \\
=& - n \int_{-\pi}^{\pi} g(x) \cos(nx) \, dx \\
=& - \pi n a_n.
\end{aligned}$$

■

Remark 4.3.3

Note that the continuity and periodicity of g can't be dropped, as can be seen by the Fourier expansion of $g(x) = x$ on $(-\pi, \pi)$.

Exercise 4.3.4 Relation Between the Fourier Series of a Function and its Derivative. Compute the Fourier series of $g(x) = x$ and $g'(x) = 1$ on $(-\pi, \pi)$, then study the relation between the two series.

Remark 4.3.5

The orthogonal family of functions $\{\sin(nx)\}_{n=1}^{\infty}$ on $[0, \pi]$ happens to be the restriction to $[0, \pi]$ of the odd functions $\sin(nx)$. For any continuous (or Riemann integrable) g on $[0, \pi]$, let g_{odd} be the odd extension of g to $[-\pi, \pi]$. Then the Fourier series of g_{odd} on $[-\pi, \pi]$ would only consist of the $\sin(nx)$ terms, and would converge to g_{odd} in the mean square on $[-\pi, \pi]$. In particular, the restriction of the Fourier series would converge to g in the mean square on $[0, \pi]$.

Note that if b_n denotes the Fourier coefficient of g_{odd} with respect to $\sin(nx)$ on $[-\pi, \pi]$, then

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} g_{\text{odd}}(x) \sin(nx)\, dx = \frac{2}{\pi} \int_{0}^{\pi} g(x) \sin(nx)\, dx.$$

The series $\sum_{n=1}^{\infty} b_n \sin(nx)$ is typically called the ***Fourier sine series*** *of g on $[0, \pi]$.*

Likewise, the family of functions $\{\cos(nx)\}_{n=0}^{\infty}$ on $[0, \pi]$ is orthogonal on $[0, \pi]$ and happens to be the restriction to $[0, \pi]$ of the even functions $\cos(nx)$. For any continuous (or Riemann integrable) g on $[0, \pi]$, let g_{even} be the even extension of g to $[-\pi, \pi]$. Then the Fourier series of g_{even} on $[-\pi, \pi]$ would only consist of the $\cos(nx)$ terms, and would converge to g_{even} in the mean square on $[-\pi, \pi]$. In particular, the restriction of the Fourier series would converge to g in the mean square on $[0, \pi]$. This series is called the ***Fourier cosine series*** *of g on $[0, \pi]$. Note that for $n \geq 1$*

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} g_{\text{even}}(x) \cos(nx)\, dx = \frac{2}{\pi} \int_{0}^{\pi} g(x) \cos(nx)\, dx,$$

while

$$a_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} g_{\text{even}}(x)\, dx = \frac{1}{\pi} \int_{0}^{\pi} g(x)\, dx.$$

Exercise 4.3.6 Fourier Sine Series. Compute the Fourier sine series of $g(x) = 1$ on $(0, \pi)$, then study the sense in which this Fourier sine series approximates $g(x)$.

Exercise 4.3.7 Fourier Cosine Series. Compute the Fourier cosine series of $g(x) = x$ on $(0, \pi)$, then study the sense in which this Fourier cosine series approximates $g(x)$.

Exercise 4.3.8 Suppose that $g(x)$ is continuous on $[-l, l]$, $g(-l) = g(l)$, and that $g'(x)$ exists except at a finite number of points and is piecewise continuous. Suppose that

$$g \sim a_0 + \sum_{n=1}^{\infty} \left[a_n \cos\left(\frac{n\pi}{l} x\right) + b_n \sin\left(\frac{n\pi}{l} x\right) \right]$$

is the Fourier series of g on $(-l,l)$. Prove that

$$\int_{-l}^{l} |g'(x)|^2\,dx = \sum_{n=1}^{\infty}\left(\frac{n\pi}{l}\right)^2 l\left[|a_n|^2+|b_n|^2\right].$$

Exercise 4.3.9 Wirtinger's inequality. Suppose that $g(x)$ is continuous on $[0,L]$, $g(0)=g(L)$, and that $g'(x)$ exists except at a finite number of points and is piecewise continuous. Prove that

$$\int_0^L |g(x)-\bar{g}|^2\,dx \le \left(\frac{L}{2\pi}\right)^2 \int_0^L |g'(x)|^2\,dx \tag{4.3.2}$$

with equality iff $g=\bar{g}+a_1\cos\left(\frac{2\pi}{L}x\right)+b_1\sin\left(\frac{2\pi}{L}x\right)$ for some constants a_1, b_1. Here $\bar{g}=L^{-1}\int_0^L g(x)\,dx$ is the average of g over $(0,L)$.

Exercise 4.3.10 Another Wirtinger's inequality. Suppose that $g(x)$ is continuous on $[0,L]$ and that $g'(x)$ exists except at a finite number of points and is piecewise continuous. Prove that

$$\int_0^L |g(x)-\bar{g}|^2\,dx \le \left(\frac{L}{\pi}\right)^2 \int_0^L |g'(x)|^2\,dx \tag{4.3.3}$$

with equality iff $g=\bar{g}+a_1\cos\left(\frac{\pi}{L}x\right)$ for some constants a_1. Here $\bar{g}=L^{-1}\int_0^L g(x)\,dx$ is the average of g over $(0,L)$.

Hint. $\{\cos\left(\frac{n\pi}{L}x\right)\}_{n=0}^{\infty}$ is a complete system of orthogonal functions on $(0,L)$ and so is $\{\sin\left(\frac{n\pi}{L}x\right)\}_{n=1}^{\infty}$. Expand g in the former and g' in the latter.

Exercise 4.3.11 Prove that a sequence of functions converges in $L^2(-l,l)$ iff the sequence of their Fourier coefficients converges in l^2.

Exercise 4.3.12 Let $M>0$ be a finite number. Consider the set S_M of continuous $g(x)$ on $[-l,l]$ such that $g'(x)$ exists except at a finite number of points and is piecewise continuous and $\left|\int_{-l}^{l} g(x)\,dx\right|, \int_{-l}^{l}|g'(x)|^2\,dx \le M$. Prove that the closure of S_M in $L^2(-l,l)$ is compact.

Hint. Use Exercise 2.8.1 and Exercise 4.3.8 to prove that any sequence in S_M has a subsequence converging in $L^2(-l,l)$.

Chapter 5

Differential Calculus of Functions of Several Variables

One main new perspective in studying the differential calculus of functions of several variables is the concept of linear approximation. The concept of partial derivatives is a generalization to the multi-dimensional setting of the concept of derivative in one-dimensional setting, but it plays a subordinate role to the concept of linear approximation.

5.1 Continuity of a Function of Several Variables

The concept of continuity of a map from a metric space to a metric space applies directly to a function of several variables---one needs to distinguish between continuity at a point, everywhere continuity, and uniform continuity. For two general metric space X and Y, there is often not much structure of the space $\mathcal{C}(X, Y)$ of maps from X to Y that are continuous everywhere (or continuous at a point) on X; for example, if $\mathbf{f}_1, \mathbf{f}_2$ are two such maps, there is usually no natural operation of $\mathbf{f}_1 + \mathbf{f}_2$ or $\mathbf{f}_1 \cdot \mathbf{f}_2$. One natural operation for this general setting is *composition*: suppose $\mathbf{f} : X \mapsto Y$ and $\mathbf{g} : Y \mapsto Z$, then $\mathbf{g} \circ \mathbf{f} : X \mapsto Z$ is defined and when $\mathbf{f}$ is continuous at $\mathbf{x}_0 \in X$ and $\mathbf{f}$ is continuous at $\mathbf{y}_0 := \mathbf{f}(\mathbf{x}_0)$, then $\mathbf{g} \circ \mathbf{f} : X \mapsto Z$ is continuous at $\mathbf{x}_0$.

However, when $Y = \mathbb{R}^m$ for some m, then $\mathbf{f}_1 + \mathbf{f}_2$ makes a natural sense; in fact, for any scalar c_1, c_2, $c_1\mathbf{f}_1 + c_2\mathbf{f}_2$ is well defined and continuous on X. This makes $\mathcal{C}(X, \mathbb{R}^m)$ a vector space.

We will just summarize a few most useful properties for this latter setting.

Let D be a subset of $\mathbb{R}^n$, $\mathbf{f} : D \mapsto \mathbb{R}^m$ be a map defined on D and $\mathbf{x}_0 \in D$. Recall that $\mathbf{f}$ is continuous at $\mathbf{x}_0$, if for any $\epsilon > 0$, there exists some $\delta > 0$ such that for any $\mathbf{x} \in D$ with $\|\mathbf{x} - \mathbf{x}_0\| < \delta$, we have $\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)\| < \epsilon$.

If we write out $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \cdots, f_m(\mathbf{x}))$, it's clear that $\mathbf{f}$ is continuous at $\mathbf{x}_0$ iff each $f_i(\mathbf{x})$, for $i = 1, \cdots, m$, is continuous at $\mathbf{x}_0$.

Remark 5.1.1

The above property does not necessarily hold when the image space is infinitely dimensional. For example, if $Y = l^2$ the space of sequences that are square summable. Suppose that $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \cdots, f_m(\mathbf{x}), \cdots) \in l^2$ and each $f_i(\mathbf{x})$, for $i = 1, \cdots, \infty$, is continuous at $\mathbf{x}_0$. Does this imply that $\mathbf{f}(\mathbf{x})$ is continuous at $\mathbf{x}_0$?

Let $\eta : \mathbb{R}^+ \mapsto \mathbb{R}^+$ be continuous such that $\eta(t) = t$ for all $0 \le t \le 1$ and $\eta(t) = 0$ for all $t \ge 2$. Then $f_m(t) = \frac{\eta(mt)}{\sqrt{m}}$ defines a continuous function on $\mathbb{R}^+$ for each m, and $\mathbf{f}(t) = (\eta(t), \cdots, \frac{\eta(mt)}{\sqrt{m}}, \cdots) \in l^2$ for each $t \in \mathbb{R}^+$, for, given any $t > 0$, $\eta(mt) = 0$ for all m such that $mt \ge 2$, so $\mathbf{f}(t)$ terminates after a finite number of terms. However, for $t = 1/N$,

$$\|\mathbf{f}(t) - \mathbf{f}(0)\|^2 = \sum_{m=1}^{\infty} \frac{\eta(mt)^2}{m} \ge \sum_{m=1}^{N} \frac{m}{N^2} \ge \frac{1}{2}$$

no matter how large N is. This shows that this $\mathbf{f}(t)$ is not continuous at $t = 0$.

Example 5.1.2 Some examples of continuous functions.

(a) The functions defining the change of coordinate from polar coordinates to rectangular coordinates are continuous for $(r, \theta) \in \mathbb{R}^+ \times [0, 2\pi]$:

$$\begin{aligned} x &= r\cos\theta \\ y &= r\sin\theta \end{aligned}$$

Part of the inverse, $r = \sqrt{x^2 + y^2}$, is defined on $\mathbb{R}^2$ and continuous there, but θ as a function of (x, y) is only defined on $\mathbb{R}^2 \backslash \{\text{a ray from } 0\}$--- one often uses the formula $\theta = \tan^{-1}(\frac{y}{x})$, but it works only for $x > 0$.

(b) The inner product function: $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbf{x} \cdot \mathbf{y} \in \mathbb{R}$ is continuous on $\mathbb{R}^n \times \mathbb{R}^n$.

Suppose $\mathbf{f}_1 : U \mapsto \mathbb{R}^n$ and $\mathbf{f}_2 : U \mapsto \mathbb{R}^n$ are two continuous functions from U to $\mathbb{R}^n$, then composition with the continuous inner product makes $\mathbf{f}_1(\mathbf{u}) \cdot \mathbf{f}_2(\mathbf{u})$ a continuous function from U to $\mathbb{R}$.

(c) Let $(X, \|\cdot\|)$ be any normed vector space, then $\mathbf{x} \mapsto \|\mathbf{x}\| \in \mathbb{R}$ is continuous from X to $\mathbb{R}$. This follows from using the triangle inequality to get

$$|\|\mathbf{x} + \mathbf{h}\| - \|\mathbf{x}\|| \le \|\mathbf{h}\|.$$

(d) Define $S(u) = u(x)^2$ for any $u \in C[a, b]$, then $u \mapsto S(u) \in C[a, b]$ is continuous in $C[a, b]$. We check that

$$|S(u+h)(x) - S(u)(x)| = |2u(x)h(x) + h(x)^2| \le 2\|u\|_{C[a,b]}\|h\|_{C[a,b]} + \|h\|^2_{C[a,b]}$$

so $\|S(u + h)(x) - S(u)(x)\|^2_{C[a,b]} \to 0$ as $\|h\|_{C[a,b]} \to 0$, proving the continuity of $S(u)$.

5.2 Linear Functions

Before we discuss linear approximation of a function from a normed vector space to a normed vector space, we first review linear functions defined from a normed vector space to a normed vector space. Even though much of the discussion makes sense in this general setting, we will mostly focus on the cases when the vector spaces are finite dimensional, in particular, when they are the usual Cartesian vector spaces. In such situations, much of this discussion is a review of some relevant concepts and properties from linear algebra. Perhaps the only new concept is the norm of a linear function (also called a linear map or a linear transformation).

5.2.1 Definition of Linear Functions and Their Matrix Representation

Definition 5.2.1

Let X, Y be two vector spaces over $\mathbb{R}$. A function $L : X \mapsto Y$ is called **linear** if

$$L(a\mathbf{x} + b\mathbf{y}) = aL(\mathbf{x}) + bL(\mathbf{y}) \quad \text{for any } \mathbf{x}, \mathbf{y} \in X, a, b \in \mathbb{R}.$$

A linear function is also called a linear map or a linear transformation .

A function $A : X \mapsto Y$ is called **affine** if there is a linear function $L : X \mapsto Y$ and a vector $\mathbf{y}_0 \in Y$ such that $A(\mathbf{x}) = \mathbf{y}_0 + L(\mathbf{x})$.

Note that $\mathbf{y}_0 = A(\mathbf{0})$, so an equivalent condition for A to be affine is that $\mathbf{x} \mapsto A(\mathbf{x}) - A(\mathbf{0})$ is linear. We often use a certain affine function $A(\mathbf{x})$ to approximate another function and in such a context we call it a linear approximation.

The most useful relation is that for any linear map $T : X \mapsto Y$ between two ***finite dimensional*** vectors spaces X and Y, once a basis $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of X and a basis $\{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ of Y are chosen, T can be represented through a matrix multiplication as follows:

$$\text{There exist coefficients } a_{ij} \text{ such that for each } 1 \le j \le n, T(\mathbf{x}_j) = \sum_{i=1}^{m} a_{ij}\mathbf{y}_i, \tag{5.2.1}$$

$$\text{then for any } \mathbf{x} \in X, \text{ there exist coefficients } c_j, 1 \le j \le n, \text{ such that } \mathbf{x} = \sum_{j=1}^{n} c_j\mathbf{x}_j,$$

thus

$$\begin{aligned} T(\mathbf{x}) &= \sum_{j=1}^{n} c_j T(\mathbf{x}_j) \\ &= \sum_{j=1}^{n} c_j \sum_{i=1}^{m} a_{ij}\mathbf{y}_i \\ &= \sum_{i=1}^{m} d_i \mathbf{y}_i \end{aligned}$$

where

$$d_i = \sum_{j=1}^{n} a_{ij}c_j, 1 \le i \le m. \tag{5.2.2}$$

In other words, the action of T in terms of the coordinates with respect to the two bases is represented through matrix multiplication by (a_{ij}).

Both (5.2.1) and (5.2.2) can be represented more cleanly using matrix notation: using

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 & \ldots & \mathbf{x}_n \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix},$$

$$\begin{bmatrix} T(\mathbf{x}_1) & \dots & T(\mathbf{x}_n) \end{bmatrix} = \begin{bmatrix} \mathbf{y}_1 & \dots & \mathbf{y}_m \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & 22 & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix},$$

we have

$$\begin{aligned} T(\mathbf{x}) &= \begin{bmatrix} T(\mathbf{x}_1) & \dots & T(\mathbf{x}_n) \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{y}_1 & \dots & \mathbf{y}_m \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & 22 & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{y}_1 & \dots & \mathbf{y}_m \end{bmatrix} \begin{bmatrix} d_1 \\ \vdots \\ d_m \end{bmatrix}, \end{aligned}$$

where

$$\begin{bmatrix} d_1 \\ \vdots \\ d_m \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & 22 & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}.$$

There is a natural addition $S + T$ of two linear maps S and T from a vector space X to a vector space Y, and a scalar multiplication cS of a linear map. This makes the set $L(X, Y)$ of linear maps from X to Y a vector space. Furthermore, when X and Y are finite dimensional, after a basis of X and a basis of Y are chosen, if S is represented by matrix A, and T is represented by matrix B, then $S + T$ is represented by matrix $A + B$.

Suppose S is a linear map from X to Y, and T is a linear map from Y to Z, then the natural composition map $T \circ S : X \mapsto Z$ is also a linear map. When X , Y and Z are all finite dimensional, and a basis has been chosen in each vector space, with A representing S and B representing T, then the matrix representation for $T \circ S$ is the matrix product BA. In fact, the matrix multiplication is defined precisely based on this natural property. We often omit the composition operator $\circ$ between S and T and write $T \circ S$ as TS.

Definition 5.2.2 Invertible Linear Map.

If $T : X \mapsto Y$ is a linear map, and there exists a linear map $S : Y \mapsto X$ such that $S \circ T = I_X$ and $T \circ S = I_Y$, namely, $S(T(\mathbf{x})) = \mathbf{x}$ for all $\mathbf{x} \in X$ and $T(S(\mathbf{y})) = \mathbf{y}$ for all $\mathbf{y} \in Y$, we say that T is an invertible linear map.

When such an S exists, it is uniquely determined. It is called the inverse of T and denoted as T^{-1}.

Exercise 5.2.3 Composition of Linear Maps. Define $T(x, y) = (x, y, x + y)$ and $S(x, y, z) = (x + y + z, x - y + z, x + z)$.

(a) Determine $S \circ T$.

(b) Find the matrix representation for T, S, and $S \circ T$ respectively in the respective standard bases.

(c) Are T or S invertible? Are they injective or surjective?

Exercise 5.2.4 Matrix Representation of the Derivative Operator. Let $\mathcal{P}_k$ denote the span of $\{1, \cos t, \sin t, \ldots, \cos(kt), \sin(kt)\}$ and define $D : \mathcal{P}_k \mapsto \mathcal{P}_k$ be the derivative operator.

(a) Find the matrix representation of D and $D \circ D$ in the given basis.

(b) Does D map the span of $\{\cos t, \sin t, \ldots, \cos(kt), \sin(kt)\}$ to itself? If so, determine whether this map is invertible.

5.2.2 Operator Norm of a Linear Map

Definition 5.2.5 Operator Norm of a Linear Map.

If X and Y are normed vector spaces, and $T : X \mapsto Y$ is a linear map, then the operator norm, also called Frobenius norm, of T is defined as

$$||T|| := \sup\{||T(\mathbf{x})||_Y : ||\mathbf{x}||_X = 1\}.$$

Equivalently,

$$||T|| := \sup\{||T(\mathbf{x})||_Y / ||\mathbf{x}||_X : \mathbf{x} \neq \mathbf{0}\}.$$

Sometimes we use the notation $||T||_{\mathcal{F}}$.

It follows that

$$||T(\mathbf{x})||_Y \leq ||T|| ||\mathbf{x}||_X \text{ for any vector } \mathbf{x},$$

and $||T||$ is the smallest number C such that

$$||T(\mathbf{x})||_Y \leq C ||\mathbf{x}||_X \text{ for any vector } \mathbf{x}.$$

In applications we often work normed vector spaces and linear maps with a finite operator norm, and when a linear map with a finite operator norm is invertible, we are interested in knowing whether its inverse also has a finite operator norm.

Remark 5.2.6

$||T||$ depends on the specific norms used on X, Y. For example,

$$I(u)(x) := \int_a^x u(t)\,dt$$

is a linear map from $X = C[a,b]$ to $Y = \{v \in C^1[a,b] : v(a) = 0\}$, where both X and Y are endowed with the $C[a,b]$ norm: $||u||_{C[a,b]} := \max_{x\in[a,b]} |u(x)|$. Then clearly $||I(u)||_{C[a,b]} \leq (b-a)||u||_{C[a,b]}$.

$I : X \mapsto Y$ here is invertible, with $I^{-1}(v) = v'(x)$ for any $v \in Y$. However, I^{-1} does not have a finite operator norm with respect to the norms we have chosen here, for, that would require the existence of some $C' > 0$ such that

$$||I^{-1}(v)||_{C[a,b]} \leq C' ||v||_{C[a,b]} \text{ for all } v \in Y.$$

But $v_k = \sin k(x-a)$ is a family in Y, with $||v_k||_{C[a,b]} = 1$, yet $||v_k'||_{C[a,b]} = |k|$, so the above inequality can't hold when $|k| \to \infty$.

If we endow Y with the $C^1[a,b]$ norm:

$$||v||_{C^1[a,b]} := \max_{x\in[a,b]} |v(x)| + \max_{x\in[a,b]} |v'(x)|,$$

then I still has a finite norm: $||I(u)||_{C^1[a,b]} \leq (b-a+1)||u||_{C[a,b]}$ for all $u \in X$, and its inverse also has a finite norm:

$$||I^{-1}(v)||_{C[a,b]} \leq ||v||_{C^1[a,b]} \text{ for all } v \in Y.$$

Remark 5.2.7

As seen from the example above, when X is not finite dimensional it is possible that a linear map from X to Y may not have a finite norm. However, when X is finite dimensional, we will prove the following proposition.

Proposition 5.2.8

Suppose that X is a finite dimensional normed vector space, then for any normed vector space Y and $T : X \mapsto Y$ a linear map from X to Y, its operator norm $||T||$ is finite.

Proof. Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a basis of X. Then any $\mathbf{x} \in X$ has coordinates $(c_1, \ldots, c_n)$ in this basis: $\mathbf{x} = \sum_{j=1}^n c_j \mathbf{x}_j$, and

$$||T(\mathbf{x})|| = ||\sum_{j=1}^n c_j T(\mathbf{x}_j)|| \leq \sum_{j=1}^n |c_j| ||T(\mathbf{x}_j)|| \leq \left(\sum_{j=1}^n ||T(\mathbf{x}_j)||^2\right)^{1/2} \left(\sum_{j=1}^n |c_j|^2\right)^{1/2}.$$

At the end of this subsection, we will prove a Lemma which implies that there exists some constant $C > 0$ such that

$$\left(\sum_{j=1}^n |c_j|^2\right)^{1/2} \leq C||\sum_{j=1}^n c_j \mathbf{x}_j|| = C||\mathbf{x}|| \text{ for all } \mathbf{x} \in X.$$

This shows that $||T||$ is finite and $||T|| \leq C \left(\sum_{j=1}^n ||T(\mathbf{x}_j)||^2\right)^{1/2}$. ■

Unless indicated otherwise, we will restrict to the situation that X is a finite dimensional normed vector space.

Suppose S and T are linear maps from X to Y. Using the property of the operator norm, we see that

$$||(T+S)(\mathbf{x})|| \leq ||T(\mathbf{x}) + S(\mathbf{x})|| \leq ||T||||\mathbf{x}|| + ||S||||\mathbf{x}|| = (||T|| + ||S||)\, ||\mathbf{x}||$$

for any vector $\mathbf{x}$, so it follows that

$$||T+S|| \leq ||T|| + ||S||.$$

It is easier to see that $||cT|| = |c|||T||$ for any scalar c. Thus the set $L(X,Y)$ of linear maps from X to Y becomes a normed vector space.

Suppose S is a linear map from X to Y, and T is a linear map from Y to Z, using the property of the operator norm, we see that

$$||TS(\mathbf{x})|| \leq ||T||||S(\mathbf{x})|| \leq ||T||||S||||\mathbf{x}|| \text{ for any vector } \mathbf{x},$$

thus $||TS|| \leq ||T||||S||$.

Remark 5.2.9

The operator norm depends on the norms on X and Y. When X and Y are Cartesian vector spaces such as $\mathbb{R}^n$, the most commonly used norm is the Euclidean norm

$$||(x_1, \dots, x_n)|| = \sqrt{\sum_{i=1}^{n} |x_i|^2}.$$

But other norms may also be used, such as

$$||(x_1, \dots, x_n)||_1 := \sum_{i=1}^{n} |x_i|,$$

or

$$||(x_1, \dots, x_n)||_\infty := \max_{1 \le i \le n} |x_i|.$$

It is usually not easy to get the precise value of the operator norm. Often one tries to give an estimate for this norm. When a linear map T is represented by a matrix A, it would make sense to estimate $||T||$ in terms of the matrix A; but one needs to be aware that T has different matrix representations depending on the choice of bases used, and any estimate of $||T||$ in terms of the matrix A has to take into account of this freedom.

When $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$ are equipped with the standard Euclidean norm, and their standard bases are used, recall that

$$T(x_1, \dots, x_n) = A \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix},$$

it follows that

$$||T(x_1, \dots, x_n)||^2 = \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix} A^{\mathrm{t}} A \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix},$$

so $||T||$ in this setting is identified as the square root of the largest eigenvalue of $A^{\mathrm{t}}A$.

Using the Cauchy-Schwarz inequality, one can easily get an estimate of the form

$$||T|| \le \sqrt{\sum_{1 \le i \le m, 1 \le j \le n} |a_{ij}|^2}.$$

The latter is actually the square mean norm on the space of matrices denoted as $\|T\|_2$.

When $m = n$ and A is an orthogonal matrix, this estimate would give $||T|| \le \sqrt{n}$ as $\sum_{1 \le j \le n} |a_{ij}|^2 = 1$ for each $1 \le i \le n$, while the above characterization, or the defining property of an orthogonal map gives $||T|| = 1$.

Exercise 5.2.10 Dependence of operator norm on vector space norm. $S(x_1, \dots, x_n) = \sum_{i=1}^{n} a_i x_i \in \mathbb{R}$ is a linear function on $\mathbb{R}^n$.

(a). Determine the operator norm of S if $\mathbb{R}^n$ is equipped with the norm $||(x_1, \dots, x_n)||_1$.

(b). Determine the operator norm of S if $\mathbb{R}^n$ is equipped with the norm $||(x_1, \dots, x_n)||_\infty$.

(c). Determine the operator norm of S if $\mathbb{R}^n$ is equipped with the norm $||(x_1, \ldots, x_n)||_p := \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ for some $1 < p < \infty$.

Remark 5.2.11

When a linear map T from X to Y has a finite norm, it is a continuous map, as

$$||T(\mathbf{x}) - T(\mathbf{y})|| = ||T(\mathbf{x} - \mathbf{y})|| \leq ||T||||\mathbf{x} - \mathbf{y}||.$$

In fact, the converse is also true. For, if $T : X \mapsto Y$ is linear and continuous at $\mathbf{0}$, then for any $\epsilon > 0$, there exists $\delta > 0$ such that $||T(\mathbf{x})|| \leq \epsilon$ whenever $||\mathbf{x}|| \leq \delta$. But for any $\mathbf{x} \neq \mathbf{0}$, $\|\frac{\delta \mathbf{x}}{||\mathbf{x}||}\| \leq \delta$, we thus have

$$||T(\frac{\delta \mathbf{x}}{||\mathbf{x}||})|| \leq \epsilon,$$

from which we conclude that

$$||T(\mathbf{x})|| \leq \frac{\epsilon}{\delta}||\mathbf{x}||.$$

When X and Y are finite dimensional, most questions about a linear map from X to Y can be formulated as a question about its matrix representation and answered that way. For example, if X and Y have the same dimension, then a a linear map T from X to Y is injective iff the null space of its matrix representation is trivial, from which one also knows that T is injective iff it is surjective.

However, when X and Y are not finite dimensional, we lose this matrix representation, and many of the conclusions or deductions in the finite dimensional setting do not work any more. For example, if $X = Y = l^2$, and L and R are the left and right shift operator respectively, then L is surjective but not injective, while R is injective but not surjective.

In the context of Fourier series, we may consider $f \mapsto (a_0, a_1, b_1, \ldots)$ as a linear map F from either $C[-\pi, \pi]$ or $\mathcal{R}[-\pi, \pi]$ to l^2, where $(a_0, a_1, b_1, \ldots)$ is the vector of Fourier coefficients of f. Then the uniqueness of the Fourier series (which is equivalent to the completeness of the sequence of standard trigonometric functions) implies that this transformation is injective. The Bessel's inequality implies that, as a linear map from $C[-\pi, \pi]$ to l^2, it has a bounded norm, as

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} |f(x)|^2 \, dx \leq (\max_{[-\pi,\pi]} |f(x)|)^2 = ||f||^2_{C[-\pi,\pi]} \text{ for all } f \in C[-\pi, \pi],$$

and

$$\begin{aligned}\|(a_0, a_1, b_1, \ldots)\|_{l^2} &\leq \left(2|a_0|^2 + \sum_{n=1}^{\infty} \left[|a_n|^2 + |b_n|^2\right]\right)^{1/2} \\ &\leq \frac{1}{\sqrt{\pi}} \left(\int_{-\pi}^{\pi} |f(x)|^2 \, dx\right)^{1/2} \\ &\leq \sqrt{2}||f||_{C[-\pi,\pi]}.\end{aligned}$$

But if we consider this linear map as from $\mathcal{R}[-\pi, \pi]$ to l^2, it does not have a bounded norm if we equip function a $f \in \mathcal{R}[-\pi, \pi]$ with the norm $\int_{-\pi}^{\pi} |f(x)| \, dx$, as for any $C > 0$, there exists $f \in \mathcal{R}[-\pi, \pi]$, such that

$$\|(a_0, a_1, b_1, \ldots)\|_{l^2} \geq \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} |f(x)|^2 \, dx\right)^{1/2} \geq C \int_{-\pi}^{\pi} |f(x)| \, dx.$$

A natural question in this context is whether F is surjective considered either as a map from $C[-\pi,\pi]$ or $\mathcal{R}[-\pi,\pi]$ to l^2. The answer to this question turns out to be negative, and it is related to whether $C[-\pi,\pi]$ or $\mathcal{R}[-\pi,\pi]$ is a ***complete*** normed space equipped with the norm $\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}|f(x)|^2\,dx\right)^{1/2}$, as the latter is also a well defined norm on either $C[-\pi,\pi]$ or $\mathcal{R}[-\pi,\pi]$.

Remark 5.2.12

The Nth Fourier series partial sum $s_N(f;x)$ of $f \in C[-\pi,\pi]$ is a linear map from $C[-\pi,\pi]$ to itself. Using the integral expression for $s_N(f;x)$, it is easy to see that its operator norm is $\frac{1}{2\pi}\int_{-\pi}^{\pi}|D_N(t)|\,dt$, which tends to ∞ as $N \to \infty$.

Fix any $x \in [-\pi,\pi]$, one may consider $f \mapsto s_N(f;x)$ as a linear map from $C[-\pi,\pi]$ to the normed vector space $\mathbb{R}$, and the norm of this transformation is also $\frac{1}{2\pi}\int_{-\pi}^{\pi}|D_N(t)|\,dt$. This fact plays a role in implying that at any designated point there are continuous functions whose Fourier series diverges there.

5.2.3 Any Two Norms on a Finite Dimensional Vector Space Are Equivalent

Lemma 5.2.13

Let X be any finite dimensional vector space over $\mathbb{R}$, and let $\mathbf{v}_1,\cdots,\mathbf{v}_n$ be a basis. For any $\mathbf{x}\in X$, let $(\mathbf{x}(1),\cdots,\mathbf{x}(n))$ be the coordinates of $\mathbf{x}$ in this basis:

$$\mathbf{x} = \sum_{k=1}^{n}\mathbf{x}(k)\mathbf{v}_k.$$

Then there exits constants $c_2 > c_1 > 0$ such that

$$c_2\left(\sum_{k=1}^{n}|\mathbf{x}(k)|^2\right)^{1/2} \ge ||\mathbf{x}|| \ge c_1\left(\sum_{k=1}^{n}|\mathbf{x}(k)|^2\right)^{1/2} \quad \forall\mathbf{x} \ \in X.$$

Proof. The general statement of the Lemma follows from the inequality above, as $\left(\sum_{k=1}^{n}|\mathbf{x}(k)|^2\right)^{1/2}$ is a concrete norm on X, and any two norms satisfy a similar inequality via their relations with this norm.

The first inequality follows from triangle inequality. For the second inequality, if it does not hold, then using homogeneity of norms, there would exist a sequence $\mathbf{x}_m$ such that

$$\left(\sum_{k=1}^{n}|\mathbf{x}_m(k)|^2\right)^{1/2} = 1, \text{ but } ||\mathbf{x}_m|| \to 0.$$

The sequence $(\mathbf{x}_m(1),\ldots,\mathbf{x}_m(n))$ is a sequence in $\mathbb{R}^n$ with unit Euclidean norm. By the Bolzano-Weierstrass Theorem, it has a convergent subsequence. Still call it $(\mathbf{x}_m(1),\ldots,\mathbf{x}_m(n))$ to simply notation, and let $(\mathbf{x}_\infty(1),\ldots,\mathbf{x}_\infty(n))$ be such that $(\mathbf{x}_m(1),\ldots,\mathbf{x}_m(n)) \to (\mathbf{x}_\infty(1),\ldots,\mathbf{x}_\infty(n))$ as $m\to\infty$. Then $\left(\sum_{k=1}^{n}|\mathbf{x}_\infty(k)|^2\right)^{1/2} = 1$. Thus $\mathbf{x}_\infty := \sum_{k=1}^{n}\mathbf{x}_\infty(k)\mathbf{v}_k \ne \mathbf{0}$.

But by the first inequality, which is already established,

$$||\mathbf{x}_\infty|| \le ||\mathbf{x}_\infty - \mathbf{x}_m|| + ||\mathbf{x}_m||$$

$$\leq c_2 \left(\sum_{k=1}^{n} |\mathbf{x}_m(k) - \mathbf{x}_\infty(k)|^2 \right)^{1/2} + ||\mathbf{x}_m||$$

Now using

$$\lim_{m\to\infty} \left(\sum_{k=1}^{n} |\mathbf{x}_m(k) - \mathbf{x}_\infty(k)|^2 \right)^{1/2} = 0 \quad \text{and} \quad \lim_{m\to\infty} ||\mathbf{x}_m|| = 0,$$

it follows that $||\mathbf{x}_\infty|| = 0$, which contradicts the property of a norm. This concludes our proof of the Lemma. ■

5.3 Differentiation

Differentiability of a function of several variables is defined in terms of a linear approximation. The notion of directional derivative and partial derivative arise when the differentiability of a function of several variables is studied along a one-dimensional line.

5.3.1 Differentiability, Directional Derivative, and Partial Derivative

Definition 5.3.1 Differentiability and Jacobian Matrix.

Suppose that X and Y are normed vector spaces, that $E \subset X$ and $\mathbf{x}$ is an interior point of E. $\mathbf{f} : E \mapsto Y$ is said to be differentiable at $\mathbf{x}$ if there exists a linear map $A : X \mapsto Y$ with finite operator norm such that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h}\|_Y}{\|\mathbf{h}\|_X} = 0. \tag{5.3.1}$$

In other words, for any $\epsilon > 0$, there exists some $\delta > 0$ such that

$$\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h}\|_Y \leq \epsilon\|\mathbf{h}\|_X \text{ for all } \mathbf{h} \text{ with } \|\mathbf{h}\| < \delta.$$

When $X = \mathbb{R}^n, Y = \mathbb{R}^m$ and $\mathbf{f}$ is differentiable at $\mathbf{x}$, the linear map $A\mathbf{h}$ can be represented as a matrix multiplication on $\mathbf{h}$ in the usual rectangular coordinates. We will denote this matrix also by A and call it the Jacobian matrix (also called Jacobian derivative or total derivative) of f at $\mathbf{x}$ and also denote it as $[D\mathbf{f}(\mathbf{x})]$.

The key here is that

- The linear approximation part, $A\mathbf{h}$, depends on $\mathbf{h}$ in a linear fashion with a finite operator norm: $||A\mathbf{h}||_Y \leq ||A||||\mathbf{h}||_X$ for all $\mathbf{h}$.
- The δ works for all $\mathbf{h}$ with $\|\mathbf{h}\| < \delta$ regardless its direction.
- If we set
$$R_\mathbf{f}(\mathbf{x};\mathbf{h}) = \mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h},$$
it is the remainder term when $\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x})$ is approximated by $A\mathbf{h}$, and (5.3.1) is equivalent to
$$\|R_\mathbf{f}(\mathbf{x};\mathbf{h})\|_Y/\|\mathbf{h}\|_X \to 0 \text{ as } \mathbf{h} \to \mathbf{0}.$$

One should check that when $\mathbf{f}$ is differentiable, then only one linear map A can satisfy the condition in the definition.

Definition 5.3.2 Directional Derivative.

Suppose $D \subset X$, $\mathbf{x}$ is an interior point of D, $\mathbf{u}$ is any non-zero vector (often taken as a unit vector). $\mathbf{f}$ is said to have directional derivative at $\mathbf{x}$ in the direction $\mathbf{u}$, if the one variable function $t \mapsto \mathbf{f}(\mathbf{x}+t\mathbf{u})$ is differentiable at $t=0$. In such a case the derivative of this one variable function at $t=0$ is called the directional derivative f at $\mathbf{x}$ in the direction $\mathbf{u}$, and is denoted as $D_{\mathbf{u}}\mathbf{f}(\mathbf{x})$.

Note that the existence of the directional derivative of $\mathbf{f}$ at $\mathbf{x}$ in the direction of $\mathbf{u}$ is defined as

$$\lim_{t\to 0}\frac{\mathbf{f}(\mathbf{x}+t\mathbf{u})-\mathbf{f}(\mathbf{x})}{t}=\mathbf{v}$$

for some vector $\mathbf{v}$. This can also be formulated as

$$\lim_{t\to 0}\|\frac{\mathbf{f}(\mathbf{x}+t\mathbf{u})-\mathbf{f}(\mathbf{x})-t\mathbf{v}}{t}\|_Y=0,$$

or equivalently, for any $\epsilon>0$, there exists some $\delta>0$ such that

$$\|\mathbf{f}(\mathbf{x}+t\mathbf{u})-\mathbf{f}(\mathbf{x})-t\mathbf{v}\|_Y<\epsilon|t| \text{ for all } t \text{ with } |t|<\delta.$$

But in this formulation there is no condition on how $\mathbf{v}$ may depend on $\mathbf{u}$ and the δ here may also depend on $\mathbf{u}$.

Remark 5.3.3

The definition of differentiability involves the norm on X and Y. But due to Lemma 5.2.13, when X is finite dimensional, it does not matter what norm to use on X.

From now on we will restrict to the case of maps between finite dimensional vector spaces, and usually take $D=\mathbb{R}^n$. When $D_{\mathbf{u}}\mathbf{f}(\mathbf{x})$ exists with $\mathbf{u}$ equal to the the standard unit vector $\mathbf{e}_j$ along the x_j coordinate, we say that $\mathbf{f}$ has **partial derivative** at $\mathbf{x}$ in the x_j variable, and denote this partial derivative as $\frac{\partial \mathbf{f}}{\partial x_j}(\mathbf{x})$. Thus $\frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_j}=D_{\mathbf{e}_j}\mathbf{f}(\mathbf{x})$. Another commonly used notation for $D_{\mathbf{e}_j}\mathbf{f}(\mathbf{x})$ is $D_j\mathbf{f}(\mathbf{x})$, or $\partial_j\mathbf{f}(\mathbf{x})$.

When $\mathbf{f}$ is differentiable at $\mathbf{x}$, then if we take any unit vector $\mathbf{u}$ and $\mathbf{h}=t\mathbf{u}$, we get

$$\lim_{\mathbf{h}\to\mathbf{0}}\frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h})-\mathbf{f}(\mathbf{x})-[D\mathbf{f}(\mathbf{x})]\mathbf{h}\|}{\|\mathbf{h}\|}=\lim_{t\to 0}\|\frac{\mathbf{f}(\mathbf{x}+t\mathbf{u})-\mathbf{f}(\mathbf{x})-t[D\mathbf{f}(\mathbf{x})]\mathbf{u}}{t}\|=0,$$

This means that the directional derivative of $\mathbf{f}$ at $\mathbf{x}$ in the direction $\mathbf{u}$ exists, and $D_{\mathbf{u}}\mathbf{f}(\mathbf{x})=[D\mathbf{f}(\mathbf{x})]\mathbf{u}$. In particular, if we take $\mathbf{u}=\mathbf{e}_j$, the standard unit vector along the x_j coordinate, we get the **partial derivative** of $\mathbf{f}$ at $\mathbf{x}$ $D_{\mathbf{e}_j}\mathbf{f}(\mathbf{x})=[D\mathbf{f}(\mathbf{x})]\mathbf{e}_j$, which is the jth column of $[D\mathbf{f}(\mathbf{x})]$.

Proposition 5.3.4

Suppose that $\mathbf{f}$ is differentiable at $\mathbf{x}$, with $[D\mathbf{f}(\mathbf{x})]$ denoting its Jacobian matrix at $\mathbf{x}$, then it is continuous at $\mathbf{x}$. Furthermore, it has directional *derivative in any direction at* $\mathbf{x}$ *and the directional derivative* $D_{\mathbf{u}}\mathbf{f}(\mathbf{x})$ *of* $\mathbf{f}$ *at* $\mathbf{x}$ *in the direction* $\mathbf{u}$ *equals* $[D\mathbf{f}(\mathbf{x})]\mathbf{u}$. *As a consequence, the* (i, j) *entry of* $[D\mathbf{f}(\mathbf{x})]$ *is* $\dfrac{\partial f_i}{\partial x_j}(\mathbf{x})$, *where* $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \cdots, f_m(\mathbf{x}))$.

Proof. The only part that we have not provided detail is to prove the continuity of $\mathbf{f}$ at $\mathbf{x}$. For any $\epsilon > 0$, there exists some $\delta > 0$ such that

$$\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h}\| \le \epsilon\|\mathbf{h}\| \text{ for all } \mathbf{h} \text{ with } \|\mathbf{h}\| < \delta.$$

Then

$$\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x})\| \le \|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h}\| + \|A\mathbf{h}\| \le \epsilon\|\mathbf{h}\| + \|A\mathbf{h}\|.$$

Using $\|A\mathbf{h}\| \le \|A\|\|\mathbf{h}\|$ we can certainly adjust $0 < \delta < 1$ to make sure that when $\|\mathbf{h}\| < \delta$, we have $\|A\mathbf{h}\| < \epsilon$, which would guarantee that $\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x})\| \le 2\epsilon$, proving the continuity of $\mathbf{f}$ at $\mathbf{x}$. ■

Question. Here are a few basic questions related to the concept of a function differentiable at a point and having partial derivatives or directional derivatives there.

- Suppose a function has partial derivatives at a point in all its coordinate directions. Does it imply that the function has directional derivatives at that point in any direction? Does it imply that the function is differentiable at that point? Does it imply that the function is continuous there?
- Suppose a function has directional derivatives at a point in any direction. Does it imply that the function is differentiable at that point? Does it imply that the function is continuous at that point?

Example 5.3.5

A function may have directional derivative at some $\mathbf{x}$ in each direction, yet can still fail to be differentiable there.

Solution. Here is a simple example.

$$f(x, y) = \begin{cases} \frac{x^2 y}{x^2+y^2} & \text{if } (x, y) \ne (0, 0), \\ 0 & \text{if } (x, y) = (0, 0). \end{cases}$$

For any direction $\mathbf{u} = (\cos\theta, \sin\theta)$, $f(t\cos\theta, t\sin\theta) = t\cos^2\theta\sin\theta$, so

$$D_{(\cos\theta, \sin\theta)} f(0, 0) = \cos^2\theta\sin\theta.$$

But it does not depend on $\mathbf{u} = (\cos\theta, \sin\theta)$ in a linear fashion, so f is not differentiable at $(0, 0)$.

A formal way of proving that this f is not differentiable at $(0, 0)$ is to argue by contradiction. If it were differentiable at $(0, 0)$, then the linear approximation must be given by $f(0, 0) + (f_x(0, 0), f_y(0, 0)) \cdot (x, y)$. But it is easy to see by definition that $f_x(0, 0) = f_y(0, 0) = 0$. Thus we would have

$$\frac{|f(x, y) - f(0, 0) - 0|}{\sqrt{x^2 + y^2}} = \frac{|x^2 y|}{(x^2 + y^2)^{3/2}} \to 0$$

as $\sqrt{x^2 + y^2} \to 0$. But that is not the case, as when $(x, y) = t(\cos\theta, \sin\theta)$, this quotient is $|\cos^2\theta\sin\theta|$, which does not tend to 0 as $\sqrt{x^2 + y^2} \to 0$.

Example 5.3.6

A function may have directional derivative at some $\mathbf{x}$ in each direction, yet can even fail to be continuous there. Here is a simple example

$$f(x,y) = \begin{cases} \frac{x^3 y}{x^6+y^2} & \text{if } (x,y) \neq (0,0), \\ 0 & \text{if } (x,y) = (0,0). \end{cases}$$

For any direction $\mathbf{u} = (\cos\theta, \sin\theta)$,

$$f(t\cos\theta, t\sin\theta) = t^2 \frac{\cos^3\theta \sin\theta}{t^4\cos^6\theta + \sin^2\theta}$$

has its derivative equal 0 at $t = 0$. However, as $t \to 0$, if we choose θ to satisfy $t^2\cos^3\theta\sin\theta$, we would get$f(t\cos\theta, t\sin\theta) = \frac{1}{2}$, which is not $\to 0$.

Example 5.3.7

A function may be differentiable at a point but may not have partial derivatives at nearby points. $f(x,y) = |xy|$ near $(0,0)$ is such an example.

In addition to the possible difference of behavior as illustrated above when the domain is more than one dimension, there may also be some differences when the function is vector-valued. If $f(x)$ is a real-valued function, differentiable on (a,b) and continuous on $[a,b]$, then the mean value theorem implies that $f(b)-f(a) = f'(c)(b-a)$ for some c between a and b. Does this property hold for vector-valued (e.g. complex-valued) functions under similar assumptions?

Exercise 5.3.8 Mean Value Theorem. Examine whether the mean value theorem holds for $\mathbf{f}(t) = (\cos t, \sin t)$ for $t \in \mathbb{R}$.

Theorem 5.3.9

Suppose that $\mathbf{f}(x)$ is a vector-valued function, differentiable on (a,b) and continuous on $[a,b]$, and that there exists some $M > 0$ such that $\|\mathbf{f}'(x)\| \le M$ for all $x \in (a,b)$. Then $\|\mathbf{f}(b) - \mathbf{f}(a)\| \le M(b-a)$.

Proof. If we assume, in addition, that $\mathbf{f}'(x)$ is continuous on (a,b), then we have an easy proof using $\mathbf{f}(b) - \mathbf{f}(a) = \int_a^b \mathbf{f}'(x)\,dx$.

For the general case, it suffices to prove that for any $\epsilon > 0$,

$$\|\mathbf{f}(c) - \mathbf{f}(a)\| \le (M+\epsilon)(c-a) + \epsilon \text{ for all } c \in [a,b]. \tag{5.3.2}$$

Fix any $\epsilon > 0$, if (5.3.2) does not hold for some c, let c^* be the infimum of the values $c \in [a,b]$ for which (5.3.2) fails. First we show that $c^* > a$. This follows from the continuity of $\mathbf{f}$ at a, as it shows that for some $\delta > 0$ (5.3.2) holds for $c \in [a, a+\delta]$. By definition of c^*, (5.3.2) holds for any $c < c^*$. By continuity of $\mathbf{f}$ at c^* (5.3.2) continues to hold at c^*. Thus $c^* < b$ under our assumption. Then using $\|\mathbf{f}'(c^*)\| \le M$, there exists some $\delta^* > 0$ such that $\|\mathbf{f}(c) - f(c^*)\| \le (M+\epsilon)(c-c^*)$ for $c \in [c^*, c^*+\delta^*]$, it then follows that, for $c \in [c^*, c^*+\delta^*]$,

$$\|\mathbf{f}(c) - \mathbf{f}(a)\| \le \|\mathbf{f}(c) - \mathbf{f}(c^*)\| + \|\mathbf{f}(c^*) - \mathbf{f}(a)\| \le (M+\epsilon)(c-c^*) + (M+\epsilon)(c^*-a) + \epsilon$$

showing that (5.3.2) continues to hold for $c \in [c^*, c^*+\delta^*]$, contradicting the definition of c^*. ■

Remark 5.3.10

For a function of several variables, in general it does not make sense to define differentiability in terms of the limit of the difference quotient

$$\frac{\mathbf{f}(\mathbf{x}+\mathbf{h})-\mathbf{f}(\mathbf{x})}{\mathbf{h}}$$

as there is no meaningful quotient operation between vectors in a general context. There are some exceptions. When $n = m = 2$, there is a well defined multiplication and quotient between vectors in $\mathbb{R}^2$ when we represent them as complex numbers. When a complex valued function $\mathbf{f}$ satisfies

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\mathbf{f}(\mathbf{x}+\mathbf{h})-\mathbf{f}(\mathbf{x})}{\mathbf{h}} = \mathbf{v} \text{ exists and is independent of how } \mathbf{h}\to\mathbf{0},$$

*it is said to have a **complex derivative**.*

This is a stronger condition than the differentiability in the linear approximation sense over vectors in $\mathbb{R}^2$ as introduced above, as that would give a function $A\mathbf{h}$ which is linear in $\mathbf{h}$ over the reals, while this condition of having a complex derivative would imply $A(i\mathbf{h}) = iA(\mathbf{h})$ and as a result $A(e^{i\theta}\mathbf{h}) = e^{i\theta}A(\mathbf{h})$ so $D_{e^{i\theta}\mathbf{u}}f(z) = e^{i\theta}D_{\mathbf{u}}f(z)$.

In terms of vector operations in $\mathbb{R}^2 \cong \mathbb{C}$, $e^{i\theta}\mathbf{u}$ corresponds to $R_\theta\mathbf{u}$, where R_θ represents rotation with respect to the origin in $\mathbb{R}^2$ of angle θ. If $\mathbf{f}$ is merely differentiable in the linear approximation sense, then

$$AR_\theta\mathbf{u} \text{ may not equal } R_\theta A\mathbf{u}.$$

In fact, a complex valued function f has a complex derivative at some z if it is differentiable in the linear approximation sense, and the linear approximation $A\mathbf{u}$ further satisfies

$$AR_\theta = R_\theta A \text{ for any rotation matrix } R_\theta. \tag{5.3.3}$$

It is easy to see that if $\mathbf{f}$ is differentiable at $\mathbf{x}$ in the linear approximation sense, then the linear approximation part $A\mathbf{h}$ is uniquely determined. It will be denoted as $\mathbf{f}'(\mathbf{x})\mathbf{h}$. It is easier to interpret this in the matrix-vector multiplication sense, with the entry in (i, j) position of $\mathbf{f}'(\mathbf{x})$ given by the partial derivative $D_jf_i(\mathbf{x}) := \frac{\partial f_i}{\partial x_j}(\mathbf{x})$.

Exercise 5.3.11 Matrix Commuting with Rotation Matrix. Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. Verify that A satisfies (5.3.3) iff there exist some real r, φ such that $A = \begin{bmatrix} r\cos\varphi & -r\sin\varphi \\ r\sin\varphi & r\cos\varphi \end{bmatrix}$.

The most important properties regarding differentiability are the chain rule and differentiability of functions with continuous partial derivatives.

5.3.2 Chain Rule

Theorem 5.3.12

Suppose that $\mathbf{f} : D \subset \mathbb{R}^n \mapsto E \subset \mathbb{R}^m$ is differentiable at $\mathbf{x}_0$, with $\mathbf{y}_0 = \mathbf{f}(\mathbf{x}_0) \in E$, and $\mathbf{g} : E \mapsto \mathbb{R}^l$ is differentiable at $\mathbf{y}_0$, then $\mathbf{g}\circ\mathbf{f}$ is differentiable at $\mathbf{x}_0$, with its Jacobian derivative $[D(\mathbf{g}\circ\mathbf{f})(\mathbf{x}_0)]$ given by the matrix product $[D\mathbf{g}(\mathbf{y}_0)][D\mathbf{f}(\mathbf{x}_0)]$.

In component notation, this is

$$D_i(g_k \circ f)(\mathbf{x}_0) = \sum_{j=1}^{m} D_j g_k \circ f(\mathbf{x}_0) D_i f_j(\mathbf{x}_0) = \sum_{j=1}^{m} D_j g_k(f(\mathbf{x}_0)) D_i f_j(\mathbf{x}_0),$$

or

$$\frac{\partial(g_k \circ f)}{\partial x_i}(\mathbf{x}_0) = \sum_{j=1}^{m} \frac{\partial g_k}{\partial y_j}(f(\mathbf{x}_0)) \frac{\partial f_j}{\partial x_i}(\mathbf{x}_0).$$

Note that we use a parenthesis as a delimiter to delineate the order of operations between composition and differentiation. Similar usage appears often such as in

$$\frac{\partial\left[f(x^2, y)\right]}{\partial x} = 2x \frac{\partial f}{\partial x}(x^2, y).$$

Proof. We first work out $\mathbf{g} \circ \mathbf{f}(\mathbf{x}) - \mathbf{g} \circ \mathbf{f}(\mathbf{x}_0)$ by using the differentiability of $\mathbf{f}$ at $\mathbf{x}_0$ and of $\mathbf{g}$ at $\mathbf{y}_0 = \mathbf{f}(\mathbf{x}_0)$:

$$\begin{aligned} \mathbf{g} \circ \mathbf{f}(\mathbf{x}) - \mathbf{g} \circ \mathbf{f}(\mathbf{x}_0) &= \mathbf{g}(\mathbf{f}(\mathbf{x})) - \mathbf{g}(\mathbf{f}(\mathbf{x}_0)) \\ &= [D\mathbf{g}(\mathbf{f}(\mathbf{x}_0))]\left[\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)\right] + \mathbf{z}(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}_0)) \end{aligned}$$

where $\|\mathbf{z}(\mathbf{y}, \mathbf{y}_0)\|/\|\mathbf{y} - \mathbf{y}_0\| \to 0$ as $\mathbf{y} \to \mathbf{y}_0$; and

$$\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0) = [D\mathbf{f}(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0) + \mathbf{w}(\mathbf{x}, \mathbf{x}_0),$$

where $\|\mathbf{w}(\mathbf{x}, \mathbf{x}_0)\|/\|\mathbf{x} - \mathbf{x}_0\| \to 0$ as $\mathbf{x} \to \mathbf{x}_0$, so we have

$$\begin{aligned} \mathbf{g} \circ \mathbf{f}(\mathbf{x}) - \mathbf{g} \circ \mathbf{f}(\mathbf{x}_0) =& [D\mathbf{g}(\mathbf{f}(\mathbf{x}_0))][D\mathbf{f}(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0) \\ &+ [D\mathbf{g}(\mathbf{f}(\mathbf{x}_0))]\mathbf{w}(\mathbf{x}, \mathbf{x}_0) + \mathbf{z}(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}_0)). \end{aligned}$$

The differentiability of $\mathbf{g} \circ \mathbf{f}(\mathbf{x})$ at $\mathbf{x}_0$ is equivalent to

$$\|[D\mathbf{g}(\mathbf{f}(\mathbf{x}_0))]\mathbf{w}(\mathbf{x}, \mathbf{x}_0) + \mathbf{z}(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}_0))\|/\|\mathbf{x} - \mathbf{x}_0\| \to 0 \text{ as } \mathbf{x} \to \mathbf{x}_0.$$

Using the property of matrix norm on $\|[D\mathbf{g}(\mathbf{f}(\mathbf{x}_0))]\mathbf{w}(\mathbf{x}, \mathbf{x}_0)\|$, we have

$$\|[D\mathbf{g}(\mathbf{f}(\mathbf{x}_0))]\mathbf{w}(\mathbf{x}, \mathbf{x}_0)\| \leq \|[D\mathbf{g}(\mathbf{f}(\mathbf{x}_0))]\|_{\mathcal{F}}\|\mathbf{w}(\mathbf{x}, \mathbf{x}_0)\|,$$

therefore

$$\|[D\mathbf{g}(\mathbf{f}(\mathbf{x}_0))]\mathbf{w}(\mathbf{x}, \mathbf{x}_0)\|/\|\mathbf{x} - \mathbf{x}_0\| \to 0 \text{ as } \mathbf{x} \to \mathbf{x}_0.$$

For $\mathbf{z}(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}_0))$, informally

$$\frac{\|\mathbf{z}(\mathbf{y}, \mathbf{y}_0)\|}{\|\mathbf{x} - \mathbf{x}_0\|} = \frac{\|\mathbf{z}(\mathbf{y}, \mathbf{y}_0)\|}{\|\mathbf{y} - \mathbf{y}_0\|} \frac{\|\mathbf{y} - \mathbf{y}_0\|}{\|\mathbf{x} - \mathbf{x}_0\|},$$

where $\mathbf{y} = \mathbf{f}(\mathbf{x})$ and $\mathbf{y}_0 = \mathbf{f}(\mathbf{x}_0)$,

$$\frac{\|\mathbf{y} - \mathbf{y}_0\|}{\|\mathbf{x} - \mathbf{x}_0\|} \text{ remains bounded as } \mathbf{x} \to \mathbf{x}_0 \text{ and } \frac{\|\mathbf{z}(\mathbf{y}, \mathbf{y}_0)\|}{\|\mathbf{y} - \mathbf{y}_0\|} \to 0 \text{ as } \mathbf{y} \to \mathbf{y}_0.$$

But this argument has a minor flaw, for $\|\mathbf{y} - \mathbf{y}_0\|$ could be 0. To fix this issue, for any $\epsilon > 0$, there exists some $\delta > 0$ such that $\|\mathbf{z}(\mathbf{y}, \mathbf{y}_0)\| \leq \epsilon\|\mathbf{y} - \mathbf{y}_0\|$ whenever $\|\mathbf{y} - \mathbf{y}_0\| < \delta$. Then using

$$\|\mathbf{y} - \mathbf{y}_0\| \leq \|\mathbf{w}(\mathbf{x}, \mathbf{x}_0) + [D\mathbf{f}(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0)\|$$

$$
\begin{aligned}
&\leq \|\mathbf{w}(\mathbf{x}, \mathbf{x}_0)\| + \|[D\mathbf{f}(\mathbf{x}_0)](\mathbf{x} - \mathbf{x}_0)\| \\
&\leq \|\mathbf{w}(\mathbf{x}, \mathbf{x}_0)\| + \|[D\mathbf{f}(\mathbf{x}_0)]\|\|\mathbf{x} - \mathbf{x}_0\|
\end{aligned}
$$

and $\frac{\|\mathbf{w}(\mathbf{x},\mathbf{x}_0)\|}{\|\mathbf{x}-\mathbf{x}_0\|} \to 0$ as $\mathbf{x} \to \mathbf{x}_0$, we can find some $\sigma > 0$ such that when $\|\mathbf{x} - \mathbf{x}_0\| < \sigma$, $\|\mathbf{w}(\mathbf{x}, \mathbf{x}_0)\| < \epsilon\|\mathbf{x} - \mathbf{x}_0\|$, so

$$\|\mathbf{y} - \mathbf{y}_0\| \leq \epsilon\|\mathbf{x} - \mathbf{x}_0\| + \|[D\mathbf{f}(\mathbf{x}_0)]\|\|\mathbf{x} - \mathbf{x}_0\| \leq (\epsilon + \|[D\mathbf{f}(\mathbf{x}_0)]\|)\,\|\mathbf{x} - \mathbf{x}_0\| < \delta.$$

Putting these together, when $\|\mathbf{x} - \mathbf{x}_0\| < \sigma$ we have

$$\|\mathbf{z}(\mathbf{y}, \mathbf{y}_0)\| \leq \epsilon\|\mathbf{y} - \mathbf{y}_0\| \leq \epsilon\,(\epsilon + \|[D\mathbf{f}(\mathbf{x}_0)]\|)\,\|\mathbf{x} - \mathbf{x}_0\|,$$

which shows the differentiability of $\mathbf{g} \circ \mathbf{f}$ at $\mathbf{x}_0$. ■

Example 5.3.13

Suppose that $\mathbf{f} : D \subset \mathbb{R}^n \mapsto E \subset \mathbb{R}^m$ and $\mathbf{g} : E \mapsto D$ are inverse of each other:

$$\mathbf{g} \circ \mathbf{f}(\mathbf{x}) = \mathbf{x} \text{ for all } \mathbf{x} \in D \text{ and } \mathbf{f} \circ \mathbf{g}(\mathbf{y}) = \mathbf{y} \text{ for all } \mathbf{y} \in E.$$

Suppose further that $\mathbf{f}$ is differentiable at $\mathbf{x}_0$, with $\mathbf{y}_0 = \mathbf{f}(\mathbf{x}_0) \in E$, and $\mathbf{g}$ is differentiable at $\mathbf{y}_0$. Then

$$[D\mathbf{g}(\mathbf{y}_0)][D\mathbf{f}(\mathbf{x}_0)] = I_{n\times n}, \quad [D\mathbf{f}(\mathbf{x}_0)][D\mathbf{g}(\mathbf{y}_0)] = I_{m\times m}.$$

As a result both $[D\mathbf{f}(\mathbf{x}_0)]$ and $[D\mathbf{g}(\mathbf{y}_0)]$ are inverse matrices of each other and $n = m$.

Exercise 5.3.14 Jacobian Matrix of Composite Function. Define $S(x, y) = (x^2 - y^2, 2xy)$ and $f(x, y) = (e^x \cos y, e^x \sin y)$. Compute the Jacobian matrix of S, f, $f \circ S$, and $S \circ f$.

Exercise 5.3.15 Chain Rule Involving Polar Coordinates. Suppose that $f(x, y)$ is differentiable for $(x, y) \in \mathbb{R}^2$. Let (r, θ) be the polar coordinates of (x, y), namely $(x, y) = P(r, \theta) = (r\cos\theta, r\sin\theta)$. Compute the Jacobian matrix of P and verify that

$$
\begin{aligned}
\frac{\partial f}{\partial r} &= \cos\theta\frac{\partial f}{\partial x} + \sin\theta\frac{\partial f}{\partial y}, \\
\frac{\partial f}{\partial \theta} &= -r\sin\theta\frac{\partial f}{\partial x} + r\cos\theta\frac{\partial f}{\partial y},
\end{aligned}
$$

In Matrix form, this is written as

$$\begin{bmatrix} \frac{\partial f}{\partial r} & \frac{\partial f}{\partial \theta} \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix} \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix}.$$

Note that we have abused the notation on the left hand side, as the function on the left hand side really represents the composition $f \circ P$ of f with P.

Note also that if we use $r^{-1}\frac{\partial f}{\partial \theta}$ instead of $\frac{\partial f}{\partial \theta}$ in the relation above we would get a simpler relation using an orthogonal matrix:

$$\begin{bmatrix} \frac{\partial f}{\partial r} & r^{-1}\frac{\partial f}{\partial \theta} \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

5.3.3 A Differentiability Criterion

Theorem 5.3.16

Suppose that $\mathbf{f} : D \subset \mathbb{R}^n \mapsto \mathbb{R}^m$ *has partial derivatives in a neighborhood of* $\mathbf{x}_0$ *and these partial derivatives are continuous at* $\mathbf{x}_0$*. Then* $\mathbf{f}$ *is differentiable at* $\mathbf{x}_0$*.*

Proof. Suppose that $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \cdots, f_m(\mathbf{x}))$. It suffices to prove that each component function $f_i(\mathbf{x})$ is differentiable at $\mathbf{x}_0$. For simplicity, we first write up a proof for the case of $n = 2$ and set $\mathbf{x}_0 = (0, 0)$.

For $\mathbf{x} = (x_1, x_2)$, we have

$$\begin{aligned} f_i(\mathbf{x}) - f_i(\mathbf{x}_0) =& f_i(x_1, x_2) - f_i(x_1, 0) + f_i(x_1, 0) - f_i(0, 0) \\ =& \frac{\partial f_i}{\partial x_2}(x_1, b)x_2 + \frac{\partial f_i}{\partial x_1}(a, 0)x_1 \\ =& \frac{\partial f_i}{\partial x_1}(0, 0)x_1 + \frac{\partial f_i}{\partial x_2}(0, 0)x_2 \\ &+ \left[\frac{\partial f_i}{\partial x_1}(a, 0) - \frac{\partial f_i}{\partial x_1}(0, 0)\right] x_1 + \left[\frac{\partial f_i}{\partial x_2}(x_1, b) - \frac{\partial f_i}{\partial x_2}(0, 0)\right] x_2 \end{aligned}$$

for some a between x_1 and 0 and some b between x_2 and 0. Using the continuity at $\mathbf{x}_0 = (0, 0)$ of the partial derivatives, for any $\epsilon > 0$, we can find some $\delta > 0$ such that whenever $\|\mathbf{x} - \mathbf{x}_0\| < \delta$, we have $|\frac{\partial f_i}{\partial x_j}(\mathbf{x}) - \frac{\partial f_i}{\partial x_j}(\mathbf{x}_0)| < \epsilon$. This implies that

$$\begin{aligned} &| \left[\frac{\partial f_i}{\partial x_1}(a, 0) - \frac{\partial f_i}{\partial x_1}(0, 0)\right] x_1 + \left[\frac{\partial f_i}{\partial x_2}(x_1, b) - \frac{\partial f_i}{\partial x_2}(0, 0)\right] x_2| \\ \leq& \epsilon\left(|x_1| + |x_2|\right) \leq \sqrt{2}\epsilon\|\mathbf{x} - \mathbf{x}_0\|, \end{aligned}$$

which shows the differentiability of $f_i(\mathbf{x})$ at $\mathbf{x}_0 = (0, 0)$. The general case can be worked out in a similar way. ■

Remark 5.3.17

The converse of the above theorem does not hold, even in one dimension. Here is a simple example:

$$f(x) = \begin{cases} x^2 \sin \frac{1}{x} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

By definition, $f'(0) = 0$*, and* $\frac{f(x) - f(0) - f'(0)x}{x} = x \sin \frac{1}{x} \to 0$ *as* $x \to 0$*, so it is differentiable at* $x = 0$*. Yet*

$$f'(x) = \begin{cases} 2x \sin \frac{1}{x} - \cos \frac{1}{x} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0, \end{cases}$$

and it's clear that $f'(x)$ *is not continuous near* $x = 0$*.*

Exercise 5.3.18 Differentiability of a norm function on $\mathbb{R}^n$. Consider the norm $||(x_1, \ldots, x_n)||_p := \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ for some $1 \leq p < \infty$ as a function on $\mathbb{R}^n$. Identify the set of points at which this function is differentiable. Repeat the exercise when the norm is $||(x_1, \ldots, x_n)||_\infty := \max_{1 \leq i \leq n} |x_i|$.

5.4 Contraction Mapping Principle

A solution to an analysis problem can often be identified and obtained as a fixed point of a certain map. The Contraction Mapping Principle gives a sufficient condition to show the existence and uniqueness (in a specified setting) of a fixed point of a certain map.

Definition 5.4.1

A map $\phi : X \mapsto X$ of a metric space (X, d) to itself is called a contraction if there exists some $c, 0 < c < 1$, such that

$$d(\phi(\mathbf{x}), \phi(\mathbf{y})) \leq c\, d(\mathbf{x}, \mathbf{y}) \quad \text{for all } \mathbf{x}, \mathbf{y} \in X.$$

Proposition 5.4.2

Suppose that $\phi : X \mapsto X$ is a contraction of X. Then it can have at most one fixed point.

Proof. Suppose that $\mathbf{x}, \mathbf{y} \in X$ are fixed points of ϕ, namely, $\phi(\mathbf{x}) = \mathbf{x}, \phi(\mathbf{y}) = \mathbf{y}$. Then

$$0 \leq d(\mathbf{x}, \mathbf{y}) = d(\phi(\mathbf{x}), \phi(\mathbf{y})) \leq c\, d(\mathbf{x}, \mathbf{y})$$

and $0 < c < 1$ imply that $d(\mathbf{x}, \mathbf{y}) = 0$. This forces $\mathbf{x} = \mathbf{y}$, proving that ϕ can have at most one fixed point. ■

Theorem 5.4.3 Fixed Point Theorem of a Contraction.

Suppose that $\phi : X \mapsto X$ is a contraction of a complete *metric space X. Then it has a unique fixed point.*

Proof. Pick any $\mathbf{x}_0 \in X$ and define $\mathbf{x}_1 = \phi(\mathbf{x}_0)$. Then define $\mathbf{x}_{k+1} = \phi(\mathbf{x}_k)$ for $k = 1, 2, \cdots$ recursively. It follows from the contraction property that

$$d(\mathbf{x}_{k+1}, \mathbf{x}_k) \leq c d(\mathbf{x}_k, \mathbf{x}_{k-1}) \leq \cdots \leq c^k d(\mathbf{x}_1, \mathbf{x}_0).$$

This then leads to $d(\mathbf{x}_{k+l}, \mathbf{x}_k) \leq \left(c^{k+l-1} + \cdots + c^k\right) d(\mathbf{x}_1, \mathbf{x}_0)$, proving that $\{\mathbf{x}_k\}$ is a Cauchy sequence in X. Since X is assumed to be complete, there exists some $\mathbf{x}$ such that $\mathbf{x}_k \to \mathbf{x}$ as $k \to \infty$. Since a contraction must be continuous, taking the limit in $\mathbf{x}_{k+1} = \phi(\mathbf{x}_k)$ gives $\mathbf{x} = \phi(\mathbf{x})$. This shows the existence of a fixed point. The uniqueness is given earlier. ■

Corollary 5.4.4

Suppose that X is a complete normed space and $A : X \mapsto Y$ is an invertible linear map from X onto another normed vector space Y with a finite operator norm for both itself and its inverse. Then any linear map $B : X \mapsto Y$ withe a finite operator norm satisfying

$$\lambda := \|A^{-1}\| \|A - B\| < 1$$

is also invertible with its inverse having a finite operator norm.

Proof. The invertibility of B is equivalent to the solvability of $B\mathbf{x} = \mathbf{y}$ for any $\mathbf{y} \in Y$. But the latter is equivalent to $A\mathbf{x} = (A - B)\mathbf{x} + \mathbf{y}$, which is equivalent to $\mathbf{x} = A^{-1}(A - B)\mathbf{x} + A^{-1}\mathbf{y}$. The solvability of the last equation is equivalent to the existence of a fixed point of

$$\phi(\mathbf{x}) = A^{-1}(A - B)\mathbf{x} + A^{-1}\mathbf{y}.$$

But under our condition, ϕ is a contraction on X, therefore, has a unique fixed point.

Furthermore, any fixed point $\mathbf{x}$ satisfies

$$\|\mathbf{x}\| \le \|A^{-1}(A - B)\mathbf{x}\| + \|A^{-1}\mathbf{y}\| \le \lambda\|\mathbf{x}\| + \|A^{-1}\mathbf{y}\|$$

from which and $\lambda < 1$ it follows that

$$\|\mathbf{x}\| \le (1 - \lambda)^{-1}\|A^{-1}\|\|\mathbf{y}\|$$

proving that B^{-1} has a finite operator norm, with $\|B^{-1}\| \le (1 - \lambda)^{-1}\|A^{-1}\|$. ■

Example 5.4.5 Existence of a Local Solution of an Initial Value Problem of an ODE.

Suppose that $f(x, t)$ is a continuous function defined on $(x, t) \in [x_0 - \delta_0, x_0 + \delta_0] \times [-\epsilon_0, \epsilon_0]$, and that there exists some $L > 0$ such that

$$|f(x, t) - f(y, t)| \le L|x - y| \text{ for all } (x, t), (y, t) \in [x_0 - \delta_0, x_0 + \delta_0] \times [-\epsilon_0, \epsilon_0].$$

Then there exists some $\epsilon, 0 < \epsilon \le \epsilon_0$ and a unique function $x(t)$ in $C^1[-\epsilon, \epsilon]$ satisfying

$$x'(t) = f(x(t), t),\ t \in [-\epsilon, \epsilon];\ x(0) = x_0. \tag{5.4.1}$$

A solution $x(t)$ to (5.4.1) is equivalent to a solution of

$$x(t) = x_0 + \int_0^t f(x(s), s)\, ds,$$

which can be regarded as a fixed point of

$$\phi(x) = x_0 + \int_0^t f(x(s), s)\, ds$$

on an appropriate space X. We will choose

$$X_\epsilon = \{x(t) \in C[-\epsilon, \epsilon] : |x(t) - x_0| \le \delta_0 \text{ for all } t \in [-\epsilon, \epsilon]\}$$

for $0 < \epsilon \le \epsilon_0$ appropriately chosen so that ϕ is a contraction on X_ϵ. Note that a fixed point $x(t)$ in X_ϵ of ϕ is automatically in $C^1[-\epsilon, \epsilon]$ and satisfied the (5.4.1).

Since $f(x, t)$ is continuous on $[x_0 - \delta_0, x_0 + \delta_0] \times [-\epsilon_0, \epsilon_0]$, there exists some $M > 0$ such that

$$|f(x, t)| \le M \text{ for all } (x, t) \in [x_0 - \delta_0, x_0 + \delta_0] \times [-\epsilon_0, \epsilon_0].$$

Then, for any $x \in X_\epsilon$,

$$|\phi(x) - x_0| \le \left|\int_0^t f(x(s), s)\, ds\right| \le M|t| \le \delta_0$$

provided that $0 < \epsilon \le \epsilon_0$ is chosen such that $M\epsilon \le \delta_0$.

Then for any $x(t), y(t) \in X_\epsilon$, we have

$$|\phi(x(t))-\phi(y(t))| \leq \left| \int_0^t (f(x(s),s) - f(y(s),s))\, ds \right| \leq L|t| \max_{s\in[-\epsilon,\epsilon]} |x(s)-y(s)|$$

for $t \in [-\epsilon, \epsilon]$. Thus if ϵ is chosen to further satisfy $L\epsilon < 1$, then ϕ becomes a contraction on X_ϵ. Note that X_ϵ is a complete metric space with the metric $d(x,y) = \max_{s\in[-\epsilon,\epsilon]} |x(s) - y(s)|$. Thus the Contraction Mapping Principle is applicable and it implies the existence and uniqueness of a fixed point of ϕ in X_ϵ.

Note that if $0 < \epsilon' < \epsilon$, then the fixed point in X_ϵ must coincide with fixed point in $X_{\epsilon'}$. Since any solution $x(t)$ of (5.4.1) must be a fixed point in $X_{\epsilon'}$ for some $\epsilon' > 0$, this shows the uniqueness in this context.

5.5 Inverse and Implicit Function Theorems

5.5.1 Inverse Function Theorem

Theorem 5.5.1

Suppose that $\mathbf{f}$ *is a differentiable mapping from a neighborhood of* $\mathbf{x}_0$ *in* $\mathbb{R}^n$ *into* $\mathbb{R}^n$*, with* $\mathbf{y}_0 = \mathbf{f}(\mathbf{x}_0)$*, and that* $D\mathbf{f}(\mathbf{x}_0)$ *is bijective and* $D\mathbf{f}(\mathbf{x})$ *is continuous at* $\mathbf{x}_0$*. Then*

1. *There is a ball* $B(\mathbf{y}_0, r)$ *in* $\mathbb{R}^n$ *and a neighborhood* U *of* $\mathbf{x}_0$ *in* $\mathbb{R}^n$ *such that* $\forall \mathbf{y} \in B(\mathbf{y}_0, r), \exists! \mathbf{x} \in U$ *satisfying* $\mathbf{f}(\mathbf{x}) = \mathbf{y}$*, and* $\mathbf{x}$ *depends on* $\mathbf{y}$ *continuously.*

2. *Denote the mapping* $\mathbf{y} \mapsto \mathbf{x}$ *by* $\mathbf{x} = \mathbf{f}^{-1}(\mathbf{y})$ *for* $\mathbf{y} \in B(\mathbf{y}_0, r)$*. Then* $\mathbf{f}^{-1}$ *is differentiable at* $\mathbf{y}_0$*, and*

$$D\mathbf{f}^{-1}(\mathbf{y}_0) = [D\mathbf{f}(\mathbf{x}_0)]^{-1}.$$

3. *If, furthermore, we assume that* $\mathbf{f}$ *is* C^1 *in this neighborhood, then* $\mathbf{f}^{-1}$ *is* C^1 *in* $B(\mathbf{y}_0, r)$ *and*

$$D\mathbf{f}^{-1}(\mathbf{y}) = [D\mathbf{f}(\mathbf{x})]^{-1} |_{x=\mathbf{f}^{-1}(\mathbf{y})}.$$

Proof. The heuristic idea is to use the differentiability of $\mathbf{f}$ to approximate $\mathbf{f}(\mathbf{x})$, when $\mathbf{x}$ is near $\mathbf{x}_0$, by the mapping $\mathbf{f}(\mathbf{x}_0) + D\mathbf{f}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$. If there were no remainder term, then solving $\mathbf{f}(\mathbf{x}) = \mathbf{y}$ would be equivalent to solving

$$\mathbf{f}(\mathbf{x}_0) + D\mathbf{f}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) = \mathbf{y},$$

which is straightforward under the assumption that $D\mathbf{f}(\mathbf{x}_0)$ is bijective. In the presence of the remainder

$$\mathbf{r}(\mathbf{x}; \mathbf{x}_0) := \mathbf{f}(\mathbf{x}) - \{\mathbf{f}(\mathbf{x}_0) + D\mathbf{f}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)\}$$

we would need to solve

$$\mathbf{r}(\mathbf{x}; \mathbf{x}_0) + \mathbf{f}(\mathbf{x}_0) + D\mathbf{f}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) = \mathbf{y}. \tag{5.5.1}$$

Using $\mathbf{y}_0 = \mathbf{f}(\mathbf{x}_0)$ and the invertibility assumption on $D\mathbf{f}(\mathbf{x}_0)$, the above is equivalent to

$$\mathbf{x} - \mathbf{x}_0 = [D\mathbf{f}(\mathbf{x}_0)]^{-1} \{\mathbf{y} - \mathbf{y}_0 - \mathbf{r}(\mathbf{x}; \mathbf{x}_0)\}.$$

In other words, we look for a fixed point $\mathbf{x}$ near $\mathbf{x}_0$ of the mapping

$$\begin{aligned}\phi(\mathbf{x}) :=& \mathbf{x}_0 + [D\mathbf{f}(\mathbf{x}_0)]^{-1}\{\mathbf{y}-\mathbf{y}_0-\mathbf{r}(\mathbf{x};\mathbf{x}_0)\}\\ =& \mathbf{x}_0 + [D\mathbf{f}(\mathbf{x}_0)]^{-1}\{\mathbf{y}-\mathbf{f}(\mathbf{x})+D\mathbf{f}(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0)\}\\ =& \mathbf{x} + [D\mathbf{f}(\mathbf{x}_0)]^{-1}\{\mathbf{y}-\mathbf{f}(\mathbf{x})\}.\end{aligned}$$

$\phi(\mathbf{x})$ is differentiable whenever $\mathbf{f}(\mathbf{x})$ is and

$$D\phi(\mathbf{x}) = I - [D\mathbf{f}(\mathbf{x}_0)]^{-1} D\mathbf{f}(\mathbf{x}) = [D\mathbf{f}(\mathbf{x}_0)]^{-1}[D\mathbf{f}(\mathbf{x}_0) - D\mathbf{f}(\mathbf{x})].$$

The continuity assumption of $D\mathbf{f}(\mathbf{x})$ at $\mathbf{x}_0$ implies the existence of some $\delta > 0$ such that

$$\| [D\mathbf{f}(\mathbf{x}_0)]^{-1}[D\mathbf{f}(\mathbf{x}_0) - D\mathbf{f}(\mathbf{x})] \| \le \frac{1}{2} \text{ for all } \mathbf{x} \text{ such that } \|\mathbf{x}-\mathbf{x}_0\| \le \delta. \tag{5.5.2}$$

By Theorem 5.3.9 it follows that

$$\|\phi(\mathbf{x}_1) - \phi(\mathbf{x}_2)\| \le \frac{1}{2}\|\mathbf{x}_1 - \mathbf{x}_2\| \tag{5.5.3}$$

for $\mathbf{x}_1, \mathbf{x}_2 \in \overline{B(\mathbf{x}_0,\delta)}$. It remains to show that one can choose an open set V containing $\mathbf{y}_0$ such that $\phi(\mathbf{x}) \in \overline{B(\mathbf{x}_0,\delta)}$ when $\mathbf{x} \in \overline{B(\mathbf{x}_0,\delta)}$ and $\mathbf{y} \in V$.

Using

$$\|\phi(\mathbf{x}_0) - \mathbf{x}_0\| \le \| [D\mathbf{f}(\mathbf{x}_0)]^{-1} \| \|\mathbf{y}-\mathbf{y}_0\| < \frac{\delta}{2},$$

if $\mathbf{y}$ satisfies $\|\mathbf{y}-\mathbf{y}_0\| < r$, where

$$r\| [D\mathbf{f}(\mathbf{x}_0)]^{-1} \| = \frac{\delta}{2}.$$

then, with (5.5.3), for any $\mathbf{x} \in \overline{B(\mathbf{x}_0,\delta)}$,

$$\|\phi(\mathbf{x}) - \mathbf{x}_0\| \le \|\phi(\mathbf{x}) - \phi(\mathbf{x}_0)\| + \|\phi(\mathbf{x}_0) - \mathbf{x}_0\| < \frac{1}{2}\|\mathbf{x}-\mathbf{x}_0\| + \frac{\delta}{2} \le \delta.$$

So if we now take $V = B(\mathbf{y}_0, r)$, then $\phi(\mathbf{x})$ would be a contraction on $\overline{B(\mathbf{x}_0,\delta)}$, thus has a unique fixed point $\mathbf{x}$ in it. In fact any fixed point $\mathbf{x}$ satisfies $\|\mathbf{x}-\mathbf{x}_0\| < \delta$. This gives a well defined inverse of $\mathbf{f}$ on $V = B(\mathbf{y}_0, r)$ with $\mathbf{f}^{-1} : V \mapsto U := \mathbf{f}^{-1}(V) \cap B(\mathbf{x}_0,\delta)$, where U is open and non-empty.

Next we prove the continuity of $\mathbf{f}^{-1}(\mathbf{y})$ for $\mathbf{y} \in B(\mathbf{y}_0, r)$. Set $\mathbf{x}_1 = \mathbf{f}^{-1}(\mathbf{y}_1), \mathbf{x}_2 = \mathbf{f}^{-1}(\mathbf{y}_2)$ for $\mathbf{y}_1, \mathbf{y}_2 \in B(\mathbf{y}_0, r)$, we examine

$$\begin{aligned}\mathbf{y}_2 - \mathbf{y}_1 =& \mathbf{f}(\mathbf{x}_2) - \mathbf{f}(\mathbf{x}_1) = (\mathbf{f}(\mathbf{x}_2) - [D\mathbf{f}(\mathbf{x}_0)]\,\mathbf{x}_2) - (\mathbf{f}(\mathbf{x}_1) - [D\mathbf{f}(\mathbf{x}_0)]\mathbf{x}_1)\\ &+ [D\mathbf{f}(\mathbf{x}_0)](\mathbf{x}_2 - \mathbf{x}_1),\end{aligned} \tag{5.5.4}$$

where we use $\mathbf{f}(\mathbf{x}) - [D\mathbf{f}(\mathbf{x}_0)]\mathbf{x}$ in place of $\mathbf{f}(\mathbf{x})$ as its derivative in $\mathbf{x}$ is $D\mathbf{f}(\mathbf{x}) - D\mathbf{f}(\mathbf{x}_0)$, which satisfies (5.5.2) when $\mathbf{x} \in B(\mathbf{x}_0,\delta)$ due to the continuity of $D\mathbf{f}(\mathbf{x})$ at $\mathbf{x}_0$. Thus

$$\| [D\mathbf{f}(\mathbf{x}_0)]^{-1}[(\mathbf{f}(\mathbf{x}_2) - [D\mathbf{f}(\mathbf{x}_0)]\,\mathbf{x}_2) - (\mathbf{f}(\mathbf{x}_1) - [D\mathbf{f}(\mathbf{x}_0)]\mathbf{x}_1)] \| \le \frac{1}{2}\|\mathbf{x}_2 - \mathbf{x}_1\|.$$

We get from (5.5.4)

$$\begin{aligned}\|\mathbf{x}_2 - \mathbf{x}_1\| \le& \| [D\mathbf{f}(\mathbf{x}_0)]^{-1}(\mathbf{y}_2 - \mathbf{y}_1) \|\\ &+ \| [D\mathbf{f}(\mathbf{x}_0)]^{-1}[(\mathbf{f}(\mathbf{x}_2) - [D\mathbf{f}(\mathbf{x}_0)]\,\mathbf{x}_2) - (\mathbf{f}(\mathbf{x}_1) - [D\mathbf{f}(\mathbf{x}_0)]\mathbf{x}_1)] \|\\ \le& \| [D\mathbf{f}(\mathbf{x}_0)]^{-1} \| \|\mathbf{y}_2 - \mathbf{y}_1\| + \frac{1}{2}\|\mathbf{x}_2 - \mathbf{x}_1\|.\end{aligned}$$

This then implies

$$\|\mathbf{x}_1 - \mathbf{x}_2\| \le 2\| \left[D\mathbf{f}(\mathbf{x}_0)\right]^{-1} \| \|\mathbf{f}(\mathbf{x}_2) - \mathbf{f}(\mathbf{x}_1)\|, \tag{5.5.5}$$

which shows the (Lipschitz) continuity of $\mathbf{f}^{-1}$ on V.

(5.5.4) can also rewritten as

$$\left[D\mathbf{f}(\mathbf{x}_0)\right]^{-1} \left[\mathbf{y}_2 - \mathbf{y}_1\right] = \mathbf{x}_2 - \mathbf{x}_1 - \left[\phi(\mathbf{x}_2) - \phi(\mathbf{x}_1)\right],$$

where ϕ is defined with respect to either $\mathbf{y}_1$ or $\mathbf{y}_2$ and satisfies (5.5.3). This would lead to the same conclusion as (5.5.5).

Finally, the relation (5.5.1) gives rise to

$$\mathbf{x} - \mathbf{x}_0 = \left[D\mathbf{f}(\mathbf{x}_0)\right]^{-1} (\mathbf{y} - \mathbf{y}_0) - \left[D\mathbf{f}(\mathbf{x}_0)\right]^{-1} \mathbf{r}(\mathbf{x}; \mathbf{x}_0).$$

Applying (5.5.5) with $\mathbf{x}, \mathbf{x}_0$ we see that, as $\|\mathbf{y} - \mathbf{y}_0\| \to 0$,

$$\frac{\| \left[D\mathbf{f}(\mathbf{x}_0)\right]^{-1} \mathbf{r}(\mathbf{x}; \mathbf{x}_0)\|}{\|\mathbf{y} - \mathbf{y}_0\|} \le \frac{2\| \left[D\mathbf{f}(\mathbf{x}_0)\right]^{-1} \|^2 \|\mathbf{r}(\mathbf{x}; \mathbf{x}_0)\|}{\|\mathbf{x} - \mathbf{x}_0\|} \to 0.$$

This shows the differentiability of $\mathbf{f}^{-1}$ at $\mathbf{y}_0$ with

$$D\mathbf{f}^{-1}(\mathbf{y}_0) = \left[D\mathbf{f}(\mathbf{x}_0)\right]^{-1}.$$

■

Remark 5.5.2

In our formulation and proof, we take V to be a ball and define $\mathbf{f}^{-1}$ on V. Since $U = \mathbf{f}^{-1}(V)$ is open and contains $\mathbf{x}_0$, there exists some $\delta' > 0$ such that $B(\mathbf{x}_0, \delta') \subset U$, then $\mathbf{f}(B(\mathbf{x}_0, \delta'))$ is an open neighborhood of $\mathbf{y}_0$ and $\mathbf{f}^{-1}$ is then a well defined inverse function of $\mathbf{f}$ on this open set.

In fact one can also prove that $\mathbf{f}$ has a well defined inverse on $B(\mathbf{x}_0, \delta)$, where δ is chosen to satisfy (5.5.2).

Exercise 5.5.3 Identify points (x_0, y_0) where the function $S(x, y) = (x^2 - y^2, 2xy)$ satisfies the conditions of the Inverse Function Theorem; at points (x_0, y_0) where the conditions of the Inverse Function Theorem are not satisfied, examine the solvability of $S(x, y) = (u, v)$ for (u, v) near $S(x_0, y_0)$.

Exercise 5.5.4 Compute the Jacobian matrix of the function $(u, v) = f(x, y) = (e^x \cos y, e^x \sin y)$ and verify that it satisfies the conditions of the Inverse Function Theorem. Determine the largest disc U around $(x, y) = (0, 0)$ on which the function f has a well-defined differentiable inverse function. Determine the largest disc V around $(u, v) = (1, 0)$ on which f^{-1} is a well-defined differentiable function. What if one drops the condition that U or V be a disc, but just requires it to be an open set?

Example 5.5.5 A geometric application of the Inverse Function Theorem.

Suppose that $\mathbf{f}(x)$ is a mapping from a neighborhood of $\mathbf{x}_0$ in $\mathbb{R}^m$ into $\mathbb{R}^n$ $(m < n)$, with $\mathbf{f}(\mathbf{x}_0) = \mathbf{y}_0 = (y_{01}, \dots, y_{0n})$, that it has continuous partial derivatives and the submatrix $[\frac{\partial f_i}{\partial x_j}], 1 \le i, j \le m$, of the the Jacobian matrix $\mathbf{f}_{\mathbf{x}}(\mathbf{x}_0)$ invertible. Such an $\mathbf{f}$ is called an **immersion** near $\mathbf{x}_0$.

Then by the Inverse Function Theorem the mapping $\mathbf{x} \mapsto (f_1(\mathbf{x}), \dots, f_m(\mathbf{x}))$

has a differentiable inverse defined in a neighborhood V of $(y_{01}, \dots, y_{0m})$. Call it Φ and let $U = \Phi(V)$. Then U is an open neighborhood of $\mathbf{x}_0$ in $\mathbb{R}^m$, and any point $\mathbf{f}(\mathbf{x})$ for $\mathbf{x} \in U$ can be represented as

$$(y_1, \dots, y_m, f_{m+1}(\Phi(y_1, \dots, y_m)), \dots, f_n(\Phi(y_1, \dots, y_m)))$$

for $(y_1, \dots, y_m) \in V$. Namely $\mathbf{f}(U)$ is represented as a graph over V with continuous partial derivatives.

Note that $\mathbf{f}$ could have been defined on a bigger domain $\mathcal{M}$, and the above discussion only says that when $\mathbf{f}$ is restricted to a small domain U near $\mathbf{x}_0$, $\mathbf{f}(U)$ is represented as a graph; it does not say that $\mathbf{f}(\mathcal{M})$ in a neighborhood of $\mathbf{f}(\mathbf{x}_0)$ is represented as a graph. A simple example is the Lemniscate of Genoro[1] given by $t \mapsto G(t) := (x, y) = (\sin t, \sin t \cos t)$ for $t \in \mathbb{R}$. Its image in $\mathbb{R}^2$ is a figure eight crossing the origin. Near $t = 0$, $x'(0) = \cos 0 = 1, y'(0) = \cos^2 0 - \sin^2 0 = 1$, so $t \mapsto x = \sin t$ has an inverse $t = \sin^{-1}(x)$ near $x = 0$, from which one gets y as a function of x, and $\frac{dy}{dx}$ can be computed via the chain rule

$$\frac{dy}{dx} = \frac{dy}{dt}\frac{dt}{dx} = \frac{dy}{dt}\Big/\frac{dx}{dt},$$

which evaluates to 1 at $t = 0$. Note that in the case here the parametric curve can be represented explicitly as $y = x\sqrt{1 - x^2}$; but this graph does not include the other branch crossing $(0, 0)$.

But we could equally apply the Inverse Function Theorem to $t \mapsto y = \sin t \cos t$ at $t = 0$ here, as $y'(0) = 1$, from which we get an inverse $t = g(y)$---we can work out an explicit form for $g(y)$ here, but we can carry on the analysis without knowing it. So the same parametric curve near $t = 0$ can also be represented as $x = \sin(g(y))$.

Exercise 5.5.6 Set $g_k(y_1, \dots, y_m) = f_k(\Phi(y_1, \dots, y_m))$ for $k = m + 1, \dots, n$ in the above setting. Use the chain rule to determine $\frac{\partial g_k}{\partial y_j}$ in terms of the $\frac{\partial f_i}{\partial x_j}, 1 \le i \le n, 1 \le j \le m$.

Exercise 5.5.7 Can the Inverse Function Theorem be applied to the Lemniscate of Genoro with respect to the x variable at $t = \pi$? If so, find $\frac{dy}{dx}$ at $(0, 0)$ where $y = g(x)$ is a graph representation for $G(\pi - \delta, \pi + \delta)$.

Can the Inverse Function Theorem be applied with respect to the y variable at $t = 0$, $t = \frac{\pi}{2}$, and $t = \pi$? If so, find $\frac{dx}{dy}$ at the corresponding point.

Can the Inverse Function Theorem be applied with respect to the x variable at $t = \frac{\pi}{2}$?

5.5.2 Implicit Function Theorem

We first introduce some notation. Suppose that $\mathbf{f}(\mathbf{x}, \mathbf{y})$ is differentiable in $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ at some $(\mathbf{x}_0, \mathbf{y}_0)$. $D\mathbf{f}(\mathbf{x}_0, \mathbf{y}_0)$ is used to represent the Jacobian derivative of $\mathbf{f}$ at $(\mathbf{x}_0, \mathbf{y}_0)$, which is a linear function on $\mathbb{R}^n \times \mathbb{R}^m$. We use $D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_0, \mathbf{y}_0)$ to represent the restriction of this linear function on $\mathbb{R}^n \times \{\mathbf{0}\}$. In other words

$$D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_0, \mathbf{y}_0)\mathbf{h} = D\mathbf{f}(\mathbf{x}_0, \mathbf{y}_0)(\mathbf{h}, 0) \quad \mathbf{h} \in \mathbb{R}^n.$$

Similarly we define $D_{\mathbf{y}}\mathbf{f}(\mathbf{x}_0, \mathbf{y}_0)$ by

$$D_{\mathbf{y}}\mathbf{f}(\mathbf{x}_0, \mathbf{y}_0)\mathbf{k} = D\mathbf{f}(\mathbf{x}_0, \mathbf{y}_0)(0, \mathbf{k}) \quad \mathbf{k} \in \mathbb{R}^m.$$

[1] www.desmos.com/calculator/skduzarzqh

If $\mathbf{f}(\mathbf{x},\mathbf{y})$ is differentiable in $\mathbf{x}$ when $\mathbf{y}$ is held fixed, we also use $D_{\mathbf{x}}\mathbf{f}(\mathbf{x},\mathbf{y})$ to denote its derivative in $\mathbf{x}$ at $(\mathbf{x},\mathbf{y})$.

This notation has a small chance of getting confused with the notation $D_{\mathbf{u}}\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)$ for the directional derivative of $\mathbf{f}$ at $(\mathbf{x}_0,\mathbf{y}_0)$ in the direction of $\mathbf{u}$, but one can usually tell the difference from the context.

Theorem 5.5.8 Implicit Function Theorem.

Let $\mathbf{f}(\mathbf{x},\mathbf{y})$ be a continuous mapping from a neighborhood $U_0 \times V_0$ of $(\mathbf{x}_0,\mathbf{y}_0)$ in $\mathbb{R}^n \times \mathbb{R}^m$ into $\mathbb{R}^n$, with $\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)=\mathbf{0}$. Assume that $\mathbf{f}(\mathbf{x},\mathbf{y})$ is differentiable in the $\mathbf{x}$ variable in this neighborhood and $D_{\mathbf{x}}\mathbf{f}(\mathbf{x},\mathbf{y})$ is continuous at $(\mathbf{x}_0,\mathbf{y}_0)$. If $D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)$ is a bijection from $\mathbb{R}^n$ onto $\mathbb{R}^n$, then

1. *There is a ball $B(\mathbf{y}_0,r) \subset V_0$ in $\mathbb{R}^m$ and a neighborhood $U \subset U_0$ of $\mathbf{x}_0$ in $\mathbb{R}^n$ such that $\forall \mathbf{y} \in B(\mathbf{y}_0,r), \exists! \mathbf{x} \in U$ satisfying $\mathbf{f}(\mathbf{x},\mathbf{y})=\mathbf{0}$, and $\mathbf{x}$ depends on $\mathbf{y}$ continuously.*

2. *Denote the mapping $\mathbf{y} \mapsto \mathbf{x}$ by $\mathbf{x}=\mathbf{u}(\mathbf{y})$. If $\mathbf{f}$ is C^1 jointly in $(\mathbf{x},\mathbf{y}) \in U \times B(\mathbf{y}_0,r)$, then $\mathbf{u}$ is C^1 in $\mathbf{y} \in B(\mathbf{y}_0,r)$ and*

$$D\mathbf{u}(\mathbf{y}) = -\left[D_{\mathbf{x}}\mathbf{f}(\mathbf{u}(\mathbf{y}),\mathbf{y})\right]^{-1} \circ D_{\mathbf{y}}\mathbf{f}(\mathbf{u}(\mathbf{y}),\mathbf{y}).$$

Proof. The heuristic for the first part is similar to that in proving the Inverse Function Theorem: for $(\mathbf{x},\mathbf{y}) \in U_0 \times V_0)$, use the linear approximation of $\mathbf{f}$ in the $\mathbf{x}$ variable to approximate $\mathbf{f}(\mathbf{x},\mathbf{y})$; more precisely,

$$\mathbf{r}(\mathbf{x};\mathbf{x}_0,\mathbf{y}) := \mathbf{f}(\mathbf{x},\mathbf{y}) - \mathbf{f}(\mathbf{x}_0,\mathbf{y}) - D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_0,\mathbf{y})(\mathbf{x}-\mathbf{x}_0).$$

Then the equation $\mathbf{f}(\mathbf{x},\mathbf{y})=\mathbf{0}$ is equivalent to

$$\mathbf{r}(\mathbf{x};\mathbf{x}_0,\mathbf{y}) = -\mathbf{f}(\mathbf{x}_0,\mathbf{y}) - D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_0,\mathbf{y})(\mathbf{x}-\mathbf{x}_0),$$

or $\mathbf{x}$ is a fixed point of the mapping

$$\begin{aligned}\phi(\mathbf{x}) :=& \mathbf{x}_0 + \left[D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_0,\mathbf{y})\right]^{-1}\left\{-\mathbf{f}(\mathbf{x}_0,\mathbf{y}) - \mathbf{r}(\mathbf{x};\mathbf{x}_0,\mathbf{y})\right\}\\ =& \mathbf{x}_0 + \left[D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_0,\mathbf{y})\right]^{-1}\left\{-\mathbf{f}(\mathbf{x},\mathbf{y}) + D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_0,\mathbf{y})(\mathbf{x}-\mathbf{x}_0)\right\}\\ =& \mathbf{x} + \left[D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_0,\mathbf{y})\right]^{-1}\left\{-\mathbf{f}(\mathbf{x},\mathbf{y})\right\}.\end{aligned}$$

Then in a similar way one can show the existence of a unique fixed point in $\overline{B(\mathbf{x}_0,\delta)}$ of ϕ when $\delta > 0$ and $r > 0$ are chosen appropriately so that ϕ satisfies (5.5.3) for $\mathbf{x} \in B(\mathbf{x}_0,\delta), \mathbf{y} \in B(\mathbf{y}_0,r)$. This shows the existence of $\mathbf{x}=\mathbf{u}(\mathbf{y})$ for $\mathbf{y} \in B(\mathbf{y}_0,r)$.

In fact, there is some flexibility in setting up ϕ. One could use a modified ϕ such as

$$\phi(\mathbf{x}) := \mathbf{x} + \left[D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)\right]^{-1}\left\{-\mathbf{f}(\mathbf{x},\mathbf{y})\right\}$$

and use its fixed point to construct $\mathbf{x}=\mathbf{u}(\mathbf{y})$.

To prove the continuity of $\mathbf{x}=\mathbf{u}(\mathbf{y})$, one takes $\mathbf{y}_1,\mathbf{y}_2 \in B(\mathbf{y}_0,r)$ and tries to use the relation

$$\begin{aligned}\mathbf{0} =& \mathbf{f}(\mathbf{u}(\mathbf{y}_2),\mathbf{y}_2) - \mathbf{f}(\mathbf{u}(\mathbf{y}_1),\mathbf{y}_1)\\ =& \mathbf{f}(\mathbf{u}(\mathbf{y}_2),\mathbf{y}_2) - \mathbf{f}(\mathbf{u}(\mathbf{y}_1),\mathbf{y}_2) + \mathbf{f}(\mathbf{u}(\mathbf{y}_1),\mathbf{y}_2) - \mathbf{f}(\mathbf{u}(\mathbf{y}_1),\mathbf{y}_1)\end{aligned}$$

and the information that $\mathbf{f}(\mathbf{u}(\mathbf{y}_1),\mathbf{y}_2) - \mathbf{f}(\mathbf{u}(\mathbf{y}_1),\mathbf{y}_1) \to \mathbf{0}$ as $\mathbf{y}_2 \to \mathbf{y}_1$ to show that $\mathbf{u}(\mathbf{y}_2) \to \mathbf{u}(\mathbf{y}_1)$.

But a standard application of the mean value theorem can only give an upper bound of $\|\mathbf{f}(\mathbf{u}(\mathbf{y}_2),\mathbf{y}_2)-\mathbf{f}(\mathbf{u}(\mathbf{y}_1),\mathbf{y}_2)\|$ in terms of $\|\mathbf{u}(\mathbf{y}_2)-\mathbf{u}(\mathbf{y}_1)\|$. Since $D_{\mathbf{x}}\mathbf{f}(\mathbf{x},\mathbf{y}_2)$ is close to $D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)$ when $\mathbf{x}\in B(\mathbf{x}_0,\delta)$ and $\mathbf{y}\in B(\mathbf{y}_0,r)$, the derivative of $\mathbf{f}(\mathbf{x},\mathbf{y}_2)-D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)\mathbf{x}$ with respect to $\mathbf{x}$ is small in the same neighorhood. In other words,

$$\mathbf{f}(\mathbf{x},\mathbf{y}_2)=D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)\mathbf{x}+[\mathbf{f}(\mathbf{x},\mathbf{y}_2)-D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)\mathbf{x}]$$

"behaves like" $D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)\mathbf{x}$ as a function of $\mathbf{x}\in B(\mathbf{x}_0,\delta)$. We implement this as

$$\begin{aligned}\mathbf{0}=&\mathbf{f}(\mathbf{u}(\mathbf{y}_2),\mathbf{y}_2)-\mathbf{f}(\mathbf{u}(\mathbf{y}_1),\mathbf{y}_1)\\ &=[D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)]\,[\mathbf{u}(\mathbf{y}_2)-\mathbf{u}(\mathbf{y}_1)]+\mathbf{f}(\mathbf{u}(\mathbf{y}_1),\mathbf{y}_2)-\mathbf{f}(\mathbf{u}(\mathbf{y}_1),\mathbf{y}_1)\\ &+\{\mathbf{f}(\mathbf{u}(\mathbf{y}_2),\mathbf{y}_2)-D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)]\mathbf{u}(\mathbf{y}_2)\}-\{\mathbf{f}(\mathbf{u}(\mathbf{y}_1),\mathbf{y}_2)-D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)]\mathbf{u}(\mathbf{y}_1)\}\end{aligned}$$

from which one gets

$$\begin{aligned}\mathbf{u}(\mathbf{y}_2)-\mathbf{u}(\mathbf{y}_1)=&-[D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)]^{-1}\{\mathbf{f}(\mathbf{u}(\mathbf{y}_1),\mathbf{y}_2)-\mathbf{f}(\mathbf{u}(\mathbf{y}_1),\mathbf{y}_1)\}\\ &+\left\{\mathbf{u}(\mathbf{y}_2)-[D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)]^{-1}\mathbf{f}(\mathbf{u}(\mathbf{y}_2),\mathbf{y}_2)\right\}\\ &-\left\{\mathbf{u}(\mathbf{y}_1)+[D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)]^{-1}\mathbf{f}(\mathbf{u}(\mathbf{y}_1),\mathbf{y}_2)\right\}\end{aligned}$$

One then uses that the $\mathbf{x}$ derivative of $\mathbf{x}\mapsto\mathbf{x}-[D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)]^{-1}\mathbf{f}(\mathbf{x},\mathbf{y}_2)$ can be made smaller than $1/2$ when $\mathbf{x}\in B(\mathbf{x}_0,\delta),\mathbf{y}_2\in B(\mathbf{y}_0,r)$ to get

$$\|\mathbf{u}(\mathbf{y}_2)-\mathbf{u}(\mathbf{y}_1)\|\le 2\|\,[D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)]^{-1}\,\|\|\mathbf{f}(\mathbf{u}(\mathbf{y}_1),\mathbf{y}_2)-\mathbf{f}(\mathbf{u}(\mathbf{y}_1),\mathbf{y}_1)\|,$$

which shows the continuity of $\mathbf{u}(\mathbf{y})$.

The differentiability of $\mathbf{u}(\mathbf{y})$ is shown in the same way as in the proof of the Inverse Function Theorem. ■

Remark 5.5.9

If, for part 1, one assumes $\mathbf{f}$ is jointly differentiable in $(\mathbf{x},\mathbf{y})$ and the derivatives are continuous at $(\mathbf{x}_0,\mathbf{y}_0)$, then one can get a simple proof by using the Inverse Function Theorem. One simply defines

$$F(\mathbf{x},\mathbf{y})=(\mathbf{f}(\mathbf{x},\mathbf{y}),\mathbf{y})\in\mathbb{R}^{n+m}.$$

Then $F(\mathbf{x}_0,\mathbf{y}_0)=(\mathbf{0},\mathbf{y}_0)$, and the Jacobian matrix at $(\mathbf{x}_0,\mathbf{y}_0)$ is an invertible $(n+m)\times(n+m)$ matrix, so there exist a neighborhood U of $(\mathbf{x}_0,\mathbf{y}_0)$, a neighborhood V of $(\mathbf{0},\mathbf{y}_0)$, an inverse G to F defined on V. When we restrict G to $(\mathbf{0},\mathbf{y})\in V$ and write out its $\mathbb{R}^n$ and $\mathbb{R}^m$components, we get

$$G(\mathbf{0},\mathbf{y})=(g(\mathbf{0},\mathbf{y}),\mathbf{y}),\quad \mathbf{f}(g(\mathbf{0},\mathbf{y}),\mathbf{y})=\mathbf{0}.$$

Remark 5.5.10

The rule for determining the derivative of the implicit function is just a form of implicit differentiation. For example, on the level set $f(x,y,z)=x^2+y^2+z^2=1$, since $f_z(x,y,z)=2z$, at any point (x,y,z) where $z\neq 0$, the Implicit Function Theorem applies to imply that z can be solved as a function of (x,y) from $x^2+y^2+z^2=1$, and

$$2x+2z\frac{\partial z}{\partial x}=0,\quad 2y+2z\frac{\partial z}{\partial y}=0$$

from which one gets $\frac{\partial z}{\partial x} = -\frac{x}{z}$*,* $\frac{\partial z}{\partial y} = -\frac{y}{z}$*.*

Our formulation allows us to study $\mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ *when* $\mathbf{f}$ *may not be differentiable in* $\mathbf{y}$*. E.g.,* $f(x, y) = x|y| + e^x$ *is differentiable in* x*,* $f(0, 0) = 1$*,* $D_x f(x, y) = |y| + e^x$ *is continuous with* $D_x f(0, 0) = 1$*, so Theorem 5.5.8 allows us to conclude that there exists a continuous* $x = g(y)$ *for* y *near* 0 *with* $g(0) = 0$ *such that* $|y|g(y) + e^{g(y)} = 1$*.*

The formulation of Theorem 5.5.8 specifies a splitting of the variables $(\mathbf{x}, \mathbf{y})$*. In applications, one often has some flexibility in making a choice of the splitting. For example, suppose that* $f(x, y, z)$ *has continuous partial derivatives, and one is interested in understanding the set* $\{(x, y, z) : f(x, y, z) = f(x_0, y_0, z_0)\}$ *for* (x, y, z) *near* (x_0, y_0, z_0)*---this is called a level set of* f*.* $f(x, y, z)$ *here is a scalar function so any application of Theorem 5.5.8 is to solve for one of the variables in terms of the remaining two.*

To be able to solve z *in terms of* (x, y)*, one needs to check whether* $D_z f(x_0, y_0, z_0)$ *as a linear map from* $\mathbb{R}$ *to* $\mathbb{R}$ *is invertible. In this case, this is equivalent to whether* $\frac{\partial f}{\partial z}(x_0, y_0, z_0) \neq 0$*. But it's possible that* $\frac{\partial f}{\partial x}(x_0, y_0, z_0) \neq 0$ *or* $\frac{\partial f}{\partial y}(x_0, y_0, z_0) \neq 0$*. Then it would mean that near* (x_0, y_0, z_0) *the set* $\{(x, y, z) : f(x, y, z) = f(x_0, y_0, z_0)\}$ *can be described as a graph of* x *in terms of* (y, z) *or a graph of* y *in terms of* (x, z)*. Thus as long as*

$$Df(x_0, y_0, z_0) = (\frac{\partial f}{\partial x}(x_0, y_0, z_0), \frac{\partial f}{\partial y}(x_0, y_0, z_0), \frac{\partial f}{\partial z}(x_0, y_0, z_0)) \neq (0, 0, 0)$$

one can apply Theorem 5.5.8 with respect to one of the variables to conclude that the level set $\{(x, y, z) : f(x, y, z) = f(x_0, y_0, z_0)\}$ *for* (x, y, z) *near* (x_0, y_0, z_0) *is given by the graph of a function of two of the variables.*

Suppose $g(x, y, z)$ *is another function with continuous partial derivatives and one is interested in understanding the set*

$$\{(x, y, z) : f(x, y, z) = f(x_0, y_0, z_0), g(x, y, z) = g(x_0, y_0, z_0)\}. \tag{5.5.6}$$

We are imposing two conditions, so an application of Theorem 5.5.8 would need to solve for two of the variables in terms of the remaining single variable. Geometrically, the set above is the intersection of a level set of f *with a level set of* g*.*

Suppose that we want to check whether the conditions of Theorem 5.5.8 hold for the (x, y) *variables, then we would need to check whether*

$$\begin{bmatrix} \frac{\partial f}{\partial x}(x_0, y_0, z_0) & \frac{\partial f}{\partial y}(x_0, y_0, z_0) \\ \frac{\partial g}{\partial x}(x_0, y_0, z_0) & \frac{\partial g}{\partial y}(x_0, y_0, z_0) \end{bmatrix}$$

is invertible. One could formulate similar criteria with respect to the other choices of a pair of variables. Note that one just needs one of the invertibility criteria to hold to apply Theorem 5.5.8 to conclude that the set in (5.5.6) near (x_0, y_0, z_0) *is a curved represented as a graph over one of the three variables. Thus a streamlined condition is that the matrix*

$$\begin{bmatrix} \frac{\partial f}{\partial x}(x_0, y_0, z_0) & \frac{\partial f}{\partial y}(x_0, y_0, z_0) & \frac{\partial f}{\partial z}(x_0, y_0, z_0) \\ \frac{\partial g}{\partial x}(x_0, y_0, z_0) & \frac{\partial g}{\partial y}(x_0, y_0, z_0) & \frac{\partial g}{\partial z}(x_0, y_0, z_0) \end{bmatrix}$$

has ranke 2*. Geometrically, this means that the two row vectors of the above matrix are linearly independent.*

Remark 5.5.11

From the structure of the proof, one can see that only the differentiability and continuity of maps, and the completeness *of the underlying space (used in the iteration scheme in the proof of the contraction mapping theorem) are used. So the Inverse Function and Implicit Function Theorems easily generalize to infinite dimensional, complete normed vector spaces, which are called* Banach spaces. *The simplest examples of such spaces are $C[0,1]$ of continuous functions on the interval $[0,1]$ with the norm $||f||_0 = \max_{x\in[0,1]} |f(x)|$.*

The Inverse Function and Implicit Function Theorems are often used to establish local solvability of solutions near a known one. In geometric context, they are often used to construct ***manifolds****. Below is an example.*

Example 5.5.12 A geometric application of the Implicit Function Theorem.

Suppose that $\mathbf{f}(x)$ is a mapping from a neighborhood of $\mathbf{x}_0 = (x_{01}, \dots, x_{0n})$ in $\mathbb{R}^n$ into $\mathbb{R}^m$ $(m < n)$, with $\mathbf{f}(\mathbf{x}_0) = \mathbf{y}_0$, that it has continuous partial derivatives and the submatrix $[\frac{\partial f_i}{\partial x_j}], 1 \le i, j \le m$, of the the Jacobian matrix $\mathbf{f}_{\mathbf{x}}(\mathbf{x}_0)$ invertible. Such an $\mathbf{f}$ is called an **submersion** near $\mathbf{x}_0$

Then by the Implicit Function Theorem there is a neighborhood V of $(x_{0(m+1)}, \dots, x_{0n})$, a neighborhood U of $(x_{01}, \dots, x_{0m})$, and a differentiable map $\Phi : V \mapsto U$ such that

$$\mathbf{f}(\Phi(x_{m+1}, \dots, x_n), x_{m+1}, \dots, x_n) = \mathbf{y}_0$$

for all $(x_{m+1}, \dots, x_n) \in V$. Furthermore, for any $(x_{m+1}, \dots, x_n) \in V$, the only solution in $U \times V$ of $\mathbf{f}(\mathbf{x}) = \mathbf{y}_0$ such that the last $(n-m)$-components of $\mathbf{x}$ is $(x_{m+1}, \dots, x_n)$ must be given by this $(\Phi(x_{m+1}, \dots, x_n), x_{m+1}, \dots, x_n)$.

For describing the set of all solutions of $\mathbf{f}(\mathbf{x}) = \mathbf{y}_0$, if $\mathbf{f}$ satisfies the assumption that its Jabobian at *any solution* of $\mathbf{f}(\mathbf{x}) = \mathbf{y}_0$ is rank m, then one can apply the Implicit Function Theorem near any solution and conclude that the set of solutions near any single solution is described by a C^1 graph over an $n-m$-dimensional ball. These are examples of what are called C^1 manifolds. The simplest such cases are when $m = 1$, where the condition on the Jacobian becomes the non-vanishing of the gradient vector $(D_1 f(\mathbf{x}_0), \dots, D_n f(\mathbf{x}_0))$, and the resulting manifold is a piece of a hypersurface.

Here is a simple example of $m = 2, n = 3$: the set

$$\{(x, y, z) : x^2 + y^2 - z^2 = 1, x - y = c\}$$

represents the intersection of the hyperboloid $\{(x, y, z) : x^2 + y^2 - z^2 = 1$ and the vertical plane $\{(x, y, z) : x - y = c\}$. We set $f(x, y, z) = x^2 + y^2 - z^2$ and $g(x, y, z) = x - y$ and check the rank of

$$\begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} & \frac{\partial f}{\partial z} \\ \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} & \frac{\partial g}{\partial z} \end{bmatrix} = \begin{bmatrix} 2x & 2y & -2z \\ 1 & -1 & 0 \end{bmatrix}.$$

Since $\det \begin{bmatrix} 2y & -2z \\ -1 & 0 \end{bmatrix} = -2z$, which is $\neq 0$ whenever $z \neq 0$, so if the intersection does not contain any point with $z = 0$, the level set can be represented in the form of $(y, z) = (g(x), h(x))$ for some differentiable g, h. The intersection has solutions with $z = 0$ only when c is the range $[-\sqrt{2}, \sqrt{2}]$. We can thus conclude that if c is not in such a range, then the intersection can be represented in the form of $(y, z) = (g(x), h(x))$ for some differentiable g, h.The derivatives can be found by implicit differentiation

$$2x + 2y\frac{\partial y}{\partial x} - 2z\frac{\partial z}{\partial x} = 0 \quad 1 - \frac{\partial y}{\partial x} = 0..$$

In the above setting one could also apply the Implicit Function Theorem to the (x, z) variables to represent the graph as functions of the y variable.

Exercise 5.5.13 Study the applicability of the Implicit Function Theorem with respect to the (x, y) variables in the level set $\{(x, y, z) : x^2 + y^2 - z^2 = 1, x - y = c\}$ for c in the range $[-\sqrt{2}, \sqrt{2}]$. When it is applicable, find $\frac{\partial x}{\partial z}$ and $\frac{\partial y}{\partial z}$.

Exercise 5.5.14 Limacon curve. The level set of $f(x, y; a) = (x^2 + y^2 - x)^2 - a^2(x^2 + y^2) = 0$ is a curve called a Limacon[2]---we treat a as a parameter.

- Identify the point(s) where this curve intersects the x axis and examine whether Theorem 5.5.8 is applicable at the point---the result may depend on the value of the parameter a.
- Note that $(0, 0)$ is a solution for any choice of a. Examine the set of solutions near $(0, 0)$ for different values of the parameter a.

Hint. Examine the plots would give some idea of the behavior to expect.

Exercise 5.5.15 Differentiability of the level set of $\|\mathbf{x}\| = c$. Consider $\|\mathbf{x}\| = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ as a function on $\mathbb{R}^n$. Identify conditions on p and $\mathbf{x}_0$ such that the level set $\{\mathbf{x} : \|\mathbf{x}\| = \|\mathbf{x}_0\|\}$ near $\mathbf{x}_0$ is the graph of a differentiable function of $(n - 1)$ variables.

Exercise 5.5.16 Foliation of a neighborhood by level surfaces. Suppose that $f(\mathbf{x})$ is continuously differentiable in a neighborhood of $\mathbf{0} \in \mathbb{R}^n$, $f(\mathbf{0}) = 0$, and $D_n f(\mathbf{0}) \neq 0$. Prove that there exist a ball $B(\mathbf{0}, r) \subset \mathbb{R}^{n-1}$, $\delta > 0$, and a continuously differentiable function $g(\mathbf{x}', t)$ defined for $(\mathbf{x}', t) \in B(\mathbf{0}, r) \times (-\delta, \delta)$ such that (a) $f(\mathbf{x}', g(\mathbf{x}', t)) = t$, (b) $g(\mathbf{0}, 0) = 0$, and (c) $V := \{(\mathbf{x}', g(\mathbf{x}', t)) : \mathbf{x}' \in B(\mathbf{0}, r)\}$ forms a neighborhood of $\mathbf{0} \in \mathbb{R}^n$.

Note that for each fixed $t \in (-\delta, \delta)$, $f = t$ on the set $\{(\mathbf{x}', g(\mathbf{x}', t)) : (\mathbf{x}', t) \in B(\mathbf{0}, r)$, so this neighborhood is **foliated** by **leaves** which are level surfaces of f.

Example 5.5.17 Differentiability of eigenvectors and eigenvalues.

Often we are interested in whether eigenvectors and eigenvalues of a matrix depend on the entries of the matrix in a continuous or differentiable way. Since any non-zero multiple of an eigenvector of a matrix is still an eigenvector, we need to impose some normalizing condition to speak of a continuously varying eigenvector.

Let A_0 be an $n \times n$ matrix and $\mathbf{x}_0 \neq \mathbf{0}$ satisfies $A_0\mathbf{x}_0 = \lambda_0\mathbf{x}_0$ for some λ_0. We normalize $\mathbf{x}_0$ such that $\|\mathbf{x}_0\| = 1$ (We use the Euclidean metric here for its desirable differentiability). Then a unit norm eigenvector and eigenvalue pair of matrix A can be formulated as to satisfy

$$F(A; \mathbf{x}, \lambda) = ((A - \lambda)\mathbf{x}, \|\mathbf{x}\|^2) = (\mathbf{0}, 1) \in \mathbb{R}^{n+1}. \tag{5.5.7}$$

Note that $F(A; \mathbf{x}, \lambda)$ has continuous partial derivatives as a function of $(A; \mathbf{x}, \lambda) \in \mathbb{R}^{n \times n} \times \mathbb{R}^n \times \mathbb{R}$. Our aim is to solve for $(\mathbf{x}, \lambda)$ as a function of A.

[2] www.desmos.com/calculator/51n6sf9wxz

Claim. Suppose that A_0 is symmetric and $\mathbf{x}_0$ is a ***simple*** eigenvector with eigenvalue λ_0, namely, the eigenspace of $A-\lambda_0 I$ is one-dimensional and spanned by $\mathbf{x}_0$. Then there exists some $\delta > 0$ and differentiable functions $Evc(A) \in \mathbb{R}^n$ and $Evl(A) \in \mathbb{R}$ defined for matrices A with $\|A - A_0\| < \delta$, such that (5.5.7) holds with $\mathbf{x} = Evc(A)$ and $\lambda = Evl(A)$.

According to the Implicit Function Theorem, we need to verify that

$$D_{(\mathbf{x},\lambda)}F(A_0, \mathbf{x}_0, \lambda_0)$$

is an invertible map of $\mathbb{R}^{n+1}$. But

$$D_{(\mathbf{x},\lambda)}F(A_0, \mathbf{x}_0, \lambda_0)(\mathbf{v}, s) = ((A_0 - \lambda_0)\mathbf{v} - s\mathbf{x}_0, 2\mathbf{x}_0 \cdot \mathbf{v}).$$

We need to establish the unique solvability in $(\mathbf{v}, s)$ of

$$(A_0 - \lambda_0 I)\mathbf{v} - s\mathbf{x}_0 = \mathbf{w} \tag{5.5.8}$$

$$2\mathbf{x}_0 \cdot \mathbf{v} = t \tag{5.5.9}$$

for any given $\mathbf{w} \in \mathbb{R}^n$, $t \in \mathbb{R}$.

Under our assumption, the matrix $A - \lambda_0 I$ has rank $(n-1)$ and $(A - \lambda_0 I)\mathbf{v} = \mathbf{c}$ has a solution iff $\mathbf{c} \cdot \mathbf{x}_0 = 0$. Thus (5.5.8) has a solution when $(s\mathbf{x}_0 + \mathbf{w}) \cdot \mathbf{x}_0 = 0$, which determines s uniquely: $s = -\mathbf{w} \cdot \mathbf{x}_0$. (5.5.9), together with (5.5.8), then determine $\mathbf{v}$ uniquely.

Note that matrix A need not be symmetric.

Question: Can either the symmetry assumption of A_0 or the simplicity assumption on the eigenspace be dropped entirely to have the same conclusions?*Hint*: Examine the behavior of eigenvectors and eigenvalues of simple 2×2 matrices.

5.6 Higher Order Derivatives and Taylor Expansion

When $\mathbf{f} : D \subset \mathbb{R}^n \mapsto \mathbb{R}^m$ is differentiable in D, then its Jacobian matrix $[D\mathbf{f}(\mathbf{x})]$ can be considered as a map from D into the vector space of $m \times n$ matrices, which can be identified with $\mathbb{R}^{m\times n}$. Thus it makes sense to consider whether $[D\mathbf{f}(\mathbf{x})]$ is differentiable in D.

If $[D\mathbf{f}(\mathbf{x})]$ is differentiable at $\mathbf{x} \in D$, then for $\mathbf{h}$ with $|\mathbf{h}|$ is small,

$$\|[D\mathbf{f}(\mathbf{x}+\mathbf{h})] - [D\mathbf{f}(\mathbf{x})] - [D\,(D\mathbf{f})](\mathbf{x})\mathbf{h}\|/\|\mathbf{h}\| \to 0, \text{ as } \mathbf{h} \to 0,$$

where $\mathbf{h} \mapsto [D\,(D\mathbf{f})](\mathbf{x})\mathbf{h}$ is a linear map from $\mathbb{R}^n$ into the vector space of $m \times n$ matrices $\mathbb{R}^{m\times n}$: if $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \cdots, f_m(\mathbf{x}))$ and $\mathbf{h} = (h_1, \cdots, h_n)$, then each of the component $\frac{\partial f_i}{\partial x_j}$ of $[D\mathbf{f}(\mathbf{x})]$ is differentiable at $\mathbf{x}$ and has directional derivatives at $\mathbf{x}$ in any direction, and

$$[D\,(D\mathbf{f})\,(\mathbf{x})]\mathbf{h} = [\sum_{k=1}^{n} h_k D_{x_k}\,(D\mathbf{f})\,(\mathbf{x})].$$

In terms of the (i, j) entry of the output matrix, it is

$$\sum_{k=1}^{n} h_k D_{x_k}\left(\frac{\partial f_i}{\partial x_j}\right)(\mathbf{x}) := \sum_{k=1}^{n} h_k \frac{\partial^2 f_i}{\partial x_k \partial x_j}(\mathbf{x}).$$

These quantities $D_{x_k}\left(\frac{\partial f_i}{\partial x_j}\right)(\mathbf{x}) = \frac{\partial^2 f_i}{\partial x_k \partial x_j}(\mathbf{x})$ are called the **second derivatives** of $f_i(\mathbf{x})$.

We will not spend energy on the more abstract concept of higher order differentials, but focus on the higher order (partial) derivatives of a scalar-valued function, where we define, say, third order derivatives of a scalar function $f(\mathbf{x})$ via

$$D^3_{x_l x_k x_j} f(\mathbf{x}) := \frac{\partial^3 f}{\partial x_l \partial x_k \partial x_j}(\mathbf{x}) = D_{x_l}\left(\frac{\partial^2 f}{\partial x_k \partial x_j}\right)(\mathbf{x}),$$

when this derivative is defined. We will often work in a setting where all the kth order partial derivatives of a function are continuous in a region, therefore any of its jth order partial derivatives, for $j \leq (k-1)$, are differentiable.

One basic question is whether the order in which to take the different mixed higher order derivatives affects the outcomes.

Example 5.6.1

Define $f(x,y) = x^y$ for $x, y > 0$. Then

$$D_x f(x,y) = yx^{y-1},\ D_y f(x,y) = x^y \ln x,$$

and

$$\begin{aligned}
\frac{\partial^2 f}{\partial y \partial x} &= \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial}{\partial y}\left(yx^{y-1}\right) = x^{y-1} + yx^{y-1}\ln x,\\
\frac{\partial^2 f}{\partial x \partial y} &= \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) = \frac{\partial}{\partial x}\left(x^y \ln x\right) = yx^{y-1}\ln x + x^{y-1},
\end{aligned}$$

So $\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial^2 f}{\partial x \partial y}$ for this function. But this property does require some conditions on the function.

Consider

$$f(x,y) = \begin{cases} xy\frac{x^2-y^2}{x^2+y^2} & (x,y) \neq (0,0) \\ 0 & (x,y) = (0,0) \end{cases}$$

Then at $(x,y) \neq (0,0)$,

$$\begin{aligned}
\frac{\partial f}{\partial x}(x,y) &= \frac{y\left(4x^2y^2 + x^4 - y^4\right)}{\left(x^2+y^2\right)^2},\\
\frac{\partial f}{\partial y}(x,y) &= \frac{-4x^3y^2 + x^5 - xy^4}{\left(x^2+y^2\right)^2},\\
\frac{\partial^2 f}{\partial y \partial x}(x,y) &= \frac{\left(x^2-y^2\right)\left(10x^2y^2 + x^4 + y^4\right)}{\left(x^2+y^2\right)^3},\\
\frac{\partial^2 f}{\partial x \partial y}(x,y) &= \frac{\left(x^2-y^2\right)\left(10x^2y^2 + x^4 + y^4\right)}{\left(x^2+y^2\right)^3},
\end{aligned}$$

so we see that

$$\frac{\partial^2 f}{\partial y \partial x}(x,y) = \frac{\partial^2 f}{\partial x \partial y}(x,y) \quad \text{when } (x,y) \neq (0,0).$$

We can also verify directly by definition that

$$\frac{\partial f}{\partial x}(0,0) = 0, \quad \frac{\partial f}{\partial y}(0,0) = 0,$$

and to compute $\frac{\partial^2 f}{\partial y \partial x}(0,0)$, we only need to examine the derivative with respect to y of $\frac{\partial f}{\partial x}(0,y) = -y$, which gives -1; while to compute $\frac{\partial^2 f}{\partial x \partial y}(0,0)$, we only need to examine the derivative with respect to x of $\frac{\partial f}{\partial y}(x,0) = x$, which gives 1. Thus

$$\frac{\partial^2 f}{\partial y \partial x}(0,0) = -1 \neq 1 = \frac{\partial^2 f}{\partial x \partial y}(0,0).$$

Theorem 5.6.2 Clairault Theorem.

Suppose that $D_i f(\mathbf{x}), D_j f(\mathbf{x}), D_{ij} f(\mathbf{x})$ exist in a neighborhood of $\mathbf{x}$ and are continuous at $\mathbf{x}$. Then $D_{ji} f(\mathbf{x})$ exists and equals $D_{ij} f(\mathbf{x})$.

Proof. We may set $\mathbf{x} = \mathbf{0}$, $i = 1, j = 2$, and $n = 2$. Then

$$\begin{aligned} D_{21} f(0,0) &= \lim_{y \to 0} \frac{D_1 f(0,y) - D_1 f(0,0)}{y} \\ &= \lim_{y \to 0} \lim_{x \to 0} \frac{f(x,y) - f(0,y) - f(x,0) + f(0,0)}{xy}. \end{aligned}$$

But applying the mean value theorem to $f(x,y) - f(0,y)$ as a function of y, we get

$$f(x,y) - f(0,y) - [f(x,0) - f(0,0)] = [D_2 f(x,y^*) - D_2 f(0,y^*)]\, y$$

for some y^* between 0 and y which may also depend on x. Applying the mean value theorem to $D_2 f(x,y^*) - D_2 f(0,y^*)$ as a function of x, we get

$$D_2 f(x,y^*) - D_2 f(0,y^*) = D_{12} f(x^*,y^*) x$$

for some x^* between 0 and x. Using the continuity of $D_{12} f(x,y)$ at $(0,0)$, it follows that

$$D_{21} f(0,0) = \lim_{y \to 0} \lim_{x \to 0} D_{12} f(x^*,y^*) = D_{12} f(0,0).$$

■

Remark 5.6.3

Note that both the formulation and proof of Clairault's theorem only involve the behavior of the function along the plane spanned by two specific coordinate directions, so it does not by itself imply the differentiability of $D_i f$. In fact, it is possible to have a function f of two variables such that $D_i f, D_{ij} f$ exist and $D_{ij} f = D_{ji} f$, yet $D_i f$ may fail to be differentiable. Here is a simple example:

$$f(x,y) = \begin{cases} \frac{x^2 y^2}{x^2 + y^2} & (x,y) \neq (0,0); \\ 0 & (x,y) = (0,0). \end{cases}$$

Fortunately in most context we will work with functions whose second order derivatives are continuous in the domain of interest so the first order derivatives are differentiable.

Definition 5.6.4

Suppose that $D \subset \mathbb{R}^n$ is open and $k \in \mathbb{N}$. We define $C^k(D)$ to be the space of functions which have j th order continuous partial derivatives in D for $1 \leq j \leq k$. When $k = 1$ we say functions in $C^1(D)$ are **continuously**

differentiable in D, and when $k > 1$, we say that functions in $C^k(D)$ are k-times continuously differentiable in D. We define $C^k(\bar{D})$ to be the space of functions in $C^k(D)$ such that each of its j th order partial derivative has a continuous extension to $\bar{D}$.

For any multi-index $\alpha = (\alpha_1, \ldots, \alpha_n)$ we denote $|\alpha| := \alpha_1 + \ldots + \alpha_n$. Note that

$$\|f\|_{C^k(\bar{D})} := \sum_{j=0}^{k} \sum_{|\alpha|=j} \max_{\mathbf{x}\in\bar{D}} |D_\alpha u(\mathbf{x})|$$

defines a norm on $C^k(\bar{D})$ and makes the latter a complete metric space.

Suppose that $f \in C^k(D)$ and $\mathbf{x} \in D$. Take any vector $\mathbf{v} \in \mathbb{R}^n$ and consider $f(\mathbf{x} + t\mathbf{v})$ as a one variable function $g(t)$ of t for t near 0. Then by the chain rule

$$g'(t) = \sum_{j=1}^{n} v_j \frac{\partial f}{\partial x_j}(\mathbf{x} + t\mathbf{v})$$

$$g''(t) = \sum_{i,j=1}^{n} v_j v_i \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x} + t\mathbf{v})$$

$$g^{(k)}(t) = \sum_{j_1,\ldots,j_k=1}^{n} v_{j_1} \cdots v_{j_k} \frac{\partial^k f}{\partial x_{j_k} \cdots \partial x_{j_1}}(\mathbf{x} + t\mathbf{v})$$

Then the one variable Taylor expansion

$$g(t) = g(0) + g'(0)t + \frac{g''(0)}{2!}t^2 + \cdots + \frac{g^{(k)}(0)}{k!}t^k + R_k(t) \tag{5.6.1}$$

gives rise to

$$\begin{aligned} f(\mathbf{x} + t\mathbf{v}) =& f(\mathbf{x}) + \sum_{j=1}^{n} t v_j \frac{\partial f}{\partial x_j}(\mathbf{x}) + \frac{t^2}{2!} \sum_{i,j=1}^{n} v_j v_i \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) \\ &+ \cdots + \frac{t^k}{k!} \sum_{j_1,\ldots,j_k=1}^{n} v_{j_1} \cdots v_{j_k} \frac{\partial^k f}{\partial x_{j_k} \cdots \partial x_{j_1}}(\mathbf{x}) + R_k(t). \end{aligned}$$

The remainder has the property that $\|R_k(t)\|/t^k \to 0$ as $t \to 0$.

To get the dependence of $R_k(t)$ in (5.6.1) on $\mathbf{v}$, we use a version of (5.6.1) with an integral remainder term:

$$g(t) = g(0) + g'(0)t + \frac{g''(0)}{2!}t^2 + \cdots + \frac{g^{(k-1)}(0)}{(k-1)!}t^{k-1} + \frac{1}{(k-1)!}\int_0^t g^{(k)}(s)(t-s)^{k-1}\,ds, \tag{5.6.2}$$

from which we find

$$R_k(t) = \frac{1}{(k-1)!}\int_0^t \left(g^{(k)}(s) - g^{(k)}(0)\right)(t-s)^{k-1}\,ds.$$

For $g(t) = f(\mathbf{x} + t\mathbf{v})$, if we make the change of variable $s = t\tau$ in the above integral, we see that $R_k(t)$ equals

$$\begin{aligned} &\frac{1}{(k-1)!}\int_0^1 \sum_{j_1,\ldots,j_k=1}^{n} t^k v_{j_1} \cdots v_{j_k} \left(\frac{\partial^k f}{\partial x_{j_k} \cdots \partial x_{j_1}}(\mathbf{x} + \tau t\mathbf{v}) - \frac{\partial^k f}{\partial x_{j_k} \cdots \partial x_{j_1}}(\mathbf{x})\right)(1-\tau)^{k-1}\,d\tau \\ =&\frac{1}{(k-1)!}\int_0^1 \sum_{j_1,\ldots,j_k=1}^{n} h_{j_1} \cdots h_{j_k} \left(\frac{\partial^k f}{\partial x_{j_k} \cdots \partial x_{j_1}}(\mathbf{x} + \tau\mathbf{h}) - \frac{\partial^k f}{\partial x_{j_k} \cdots \partial x_{j_1}}(\mathbf{x})\right)(1-\tau)^{k-1}\,d\tau \end{aligned}$$

so $R_k(t)$ is actually a function of $\mathbf{x}$ and $\mathbf{h} = t\mathbf{v}$.

Using the continuity of $\frac{\partial^k f}{\partial x_{j_k} \cdots \partial x_{j_1}}$ at $\mathbf{x}$, we find that, for any $\epsilon > 0$, there exists some $\delta > 0$ such that

$$\left\| \frac{\partial^k f}{\partial x_{j_k} \cdots \partial x_{j_1}}(\mathbf{x}+\mathbf{h}) - \frac{\partial^k f}{\partial x_{j_k} \cdots \partial x_{j_1}}(\mathbf{x}) \right\| < \epsilon$$

for all $\mathbf{h}$ with $\|\mathbf{h}\| < \delta$. Thus when $0 \le \|\mathbf{h}\| < \delta$, we have

$$\begin{aligned} |R_k(t)| \le & \frac{\epsilon}{(k-1)!} \int_0^1 \sum_{j_1,\ldots,j_k=1}^{n} |h_{j_1}| \cdots |h_{j_k}| (1-s)^{k-1}\, ds \\ \le & \frac{C(n,k)\epsilon \|\mathbf{h}\|^k}{k!} \end{aligned}$$

where we have used $\sum_{j_1,\ldots,j_k=1}^{n} |h_{j_1}| \cdots |h_{j_k}| \le C(n,k)\|\mathbf{h}\|^k$.

We summarize this as

Theorem 5.6.5 Taylor Expansion.

Suppose that $f \in C^k(D)$ and $\mathbf{x} \in D \subset \mathbb{R}^n$. Then the kth order Taylor expansion of f at $\mathbf{x}$, $T_k(f,\mathbf{x})(\mathbf{h})$, defined as

$$f(\mathbf{x}) + \sum_{j=1}^{n} h_j \frac{\partial f}{\partial x_j}(\mathbf{x}) + \frac{1}{2!} \sum_{i,j=1}^{n} h_j h_i \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) + \cdots + \frac{1}{k!} \sum_{j_1,\ldots,j_k=1}^{n} h_{j_1} \cdots h_{j_k} \frac{\partial^k f}{\partial x_{j_k} \cdots \partial x_{j_1}}(\mathbf{x})$$

satisfies

$$\|f(\mathbf{x}+\mathbf{h}) - T_k(f,\mathbf{x})(\mathbf{h})\| / \|\mathbf{h}\|^k \to 0 \text{ as } \mathbf{h} \to \mathbf{0}. \tag{5.6.3}$$

Furthermore, under the assumption here, for any subdomain D' of D such that the closure of D' is compact and any $\epsilon > 0$, there exists $\delta > 0$ such that

$$\|f(\mathbf{x}+\mathbf{h}) - T_k(f,\mathbf{x})(\mathbf{h})\| \le \epsilon \|\mathbf{h}\|^k \text{ for all } \mathbf{x} \in D', \|\mathbf{h}\| < \delta.$$

Remark 5.6.6

(5.6.3) can also be established under the weaker assumption that all partial derivatives of f of order up to $k-1$ are defined and continuous in a neighborhood of $\mathbf{x}$, and all all partial derivatives of f of order $k-1$ are differentiable at $\mathbf{x}$.

$T_k(f,\mathbf{x})(\mathbf{h})$ is a polynomial in $\mathbf{h}$ of degree at most k. There are contexts where one works with a function f with the property that at some $\mathbf{x}$, there exists a polynomial in $\mathbf{h}$ of degree at most k, $P_k(\mathbf{x};\mathbf{h})$ such that

$$\|f(\mathbf{x}+\mathbf{h}) - P_k(\mathbf{x};\mathbf{h})\| / \|\mathbf{h}\|^k \to 0 \text{ as } \mathbf{h} \to \mathbf{0}. \tag{5.6.4}$$

When this holds, such a $P_k(\mathbf{x};\mathbf{h})$ is unique and f is differentiable at $\mathbf{x}$ with $f(\mathbf{x}) = P_k(\mathbf{x};\mathbf{0})$, $D_{\mathbf{x}} f(\mathbf{x}) = D_{\mathbf{h}} P_k(\mathbf{x};\mathbf{0})$, but f may not have derivatives at all nearby points.

The expansion (5.6.3) is used often when $k = 2$, where we can write

$$\sum_{i,j=1}^{n} v_j v_i \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) = \mathbf{v}^{\mathrm{t}} [D^2 f(\mathbf{x})] \mathbf{v}$$

with $[D^2 f(\mathbf{x})]$ denoting the Hessian matrix of f at $\mathbf{x}$ with entries $\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x})$, and $\mathbf{v}^{\mathrm{t}}$ denoting the transpose of $\mathbf{v}$. If $\mathbf{x}$ is an interior minimum of f, then for any vector $\mathbf{v}$, the one variable function $f(\mathbf{x}+t\mathbf{v})$ has $t=0$ as an interior minimum. Therefore

$$g''(0) = \sum_{i,j=1}^{n} v_j v_i \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) \geq 0.$$

This then implies that the Hessian matrix $[D^2 f(\mathbf{x})]$ is non-negative definite. Conversely, if $\mathbf{x}$ is an interior critical point of a twice continuously differentiable f, namely, $D_i f(\mathbf{x}) = 0$ for all $i = 1, \ldots, n$, and $[D^2 f(\mathbf{x})]$ is positive definite, then the Taylor expansion of order 2 above would show that $\mathbf{x}$ is a local minimum of f.

Exercise 5.6.7 Prove (5.6.3) under the assumption that all partial derivatives of f of order up to $k-1$ are defined and continuous in a neighborhood of $\mathbf{x}$, and all all partial derivatives of f of order $(k-1)$ are differentiable at $\mathbf{x}$.Use (5.6.2) with the integral remainder term at order $k-1$ and use the differentiability of the order $(k-1)$ partial derivatives of f at $\mathbf{x}$, which appear in $g^{k-1}(s)$ in the integral remainder term.

Exercise 5.6.8 Prove that, if (5.6.4) holds for some $P_k(\mathbf{x};\mathbf{h})$, then it is unique.

Exercise 5.6.9 Construct an example of a function such that (5.6.4) holds for some $P_k(\mathbf{x};\mathbf{h})$, but f fails to have derivatives at a sequence $\mathbf{x}_m \to \mathbf{x}$.

5.7 Differentiation of Integrals

Suppose that $\phi(x,t)$ is differentiable with respect to t when each x is held fixed. We are interested in knowing under what conditions the following holds

$$D_t \int_a^b \phi(x,t)\, dx = \int_a^b D_t \phi(x,t)\, dx. \tag{5.7.1}$$

The question is really a case of Theorem 2.2.7, or our extension Theorem 2.3.2. For it is really asking whether

$$\lim_{h_n \to 0} \int_a^b \frac{\phi(x,t+h_n) - \phi(x,t)}{h_n}\, dx = \int_a^b D_t \phi(x,t)\, dx.$$

The condition in Theorem 2.2.7 requires that $[a,b]$ be a finite interval and $\frac{\phi(x,t+h_n)-\phi(x,t)}{h_n} \to D_t\phi(x,t)$ uniformly over $x \in [a,b]$ as $n \to \infty$. Since the mean value theorem gives us $\frac{\phi(x,t+h_n)-\phi(x,t)}{h_n} = D_t\phi(x,t+\theta h_n)$ for some $0<\theta<1$ depending on x, t, h_n, a sufficient condition for this uniform convergence is the following condition: For any $\epsilon > 0$ there exists a $\delta > 0$ such that for all $x \in [a,b], s \in (t-\delta, t+\delta)$

$$|D_t\phi(x,s) - D_t\phi(x,t)| < \epsilon.$$

When the uniform integrability fails, differentiation under the integral may not hold.

Example 5.7.1 Differentiability of the integral $\int_0^1 \frac{t^2 x}{(t+x)^3}\, dx$.

The integrand $f(x,t) = \frac{t^2 x}{(t+x)^3}$ is continuously differentiable in the first quadrant. For any $0 < x \leq 1$, $f(x,t) \to 0$ as $t \to 0+$, so we may extend $f(x,0)=0$ to make the resulting function continuous for $x > 0, t \geq 0$ (is it continuous at $(x,t)=(0,0)$?). We can consider either $\frac{f(x,t)-f(x,0)}{t}$ as $t\to 0+$, or the limit of $D_tf(x,t)=\frac{2tx(t+x)-3t^2x}{(t+x)^4}$ as $t\to 0+$.

In the particular example here,

$$\int_0^1 \frac{f(x,t)-f(x,0)}{t}\,dx=\int_0^1\frac{tx}{(t+x)^3}\,dx=\int_0^{1/t}\frac{y}{(1+y)^3}\,dy$$

which tends to $\int_0^\infty \frac{y}{(1+y)^3}\,dy>0$ as $t\to 0+$. On the other hand, $\frac{f(x,t)-f(x,0)}{t}\to 0$ for every $0<x\le 1$ as $t\to 0+$. It is instructive to examine how the uniform integrability has failed here.

This example is constructed based on an earlier example when examining

$$\lim_{n\to\infty}\int_0^1\frac{n^2x}{(1+nx)^3}\,dx=\lim_{n\to\infty}\int_0^1\frac{\frac{1}{n}x}{(\frac{1}{n}+x)^2}\,dx.$$

We simply replace $\frac{1}{n}$ by $t>0$ to get this example.

When the interval of integration is unbounded, even if we can make $|\frac{\phi(x,t+h)-\phi(x,t)}{h}-\phi_t(x,t)|<\epsilon$, uniformly for all x in the interval of integration and h with $|h|<\delta$, this may not guarantee that $\left|\int_a^b\left(\frac{\phi(x,t+h)-\phi(x,t)}{h}-\phi_t(x,t)\right)dx\right|$ is small when a or b is infinite.

For definiteness of exposition, let's assume a to be finite and $b=\infty$. We will also use $f(x,t)$ in place of $\phi(x,t)$. Our strategy is to break the integral into an integral on a finite integral, which we can handle by the earlier method, and the tail part, which will be made small by some reasonable assumption. More specifically, we assume that the improper integral $\int_a^\infty D_tf(x,s)\,dx$ converges *uniformly with respect to s in some neighborhood I* of t, namely,

> For any $\epsilon>0$, there exists L such that for any $N>M\ge L$, $|\int_M^N D_tf(x,s)\,dx|<\epsilon/4$ for all $s\in I$.

Theorem 5.7.2

Suppose that there exists a neighborhood I of t such that $f(x,s)$ and $D_tf(x,s)$ are continuous in $[a,\infty)\times I$, that the improper integral $\int_a^\infty f(x,s)\,dx$ converges for $s\in I$, and that the improper integral $\int_a^\infty D_tf(x,s)\,dx$ converges uniformly with respect to s in I. Then $\int_a^\infty f(x,t)\,dx$ is differentiable for $t\in I$, and

$$\frac{d}{dt}\left(\int_a^\infty f(x,t)\,dx\right)=\int_a^\infty D_tf(x,t)\,dx$$

for $t\in I$.

Proof. First we break up the integrals into parts and estimate them separately

$$\begin{aligned}&\left|\int_a^\infty\left(\frac{f(x,t+h)-f(x,t)}{h}-D_tf(x,t)\right)dx\right|\\ \le&\left|\int_a^L\left(\frac{f(x,t+h)-f(x,t)}{h}-D_tf(x,t)\right)dx\right|+\left|\int_L^\infty\frac{f(x,t+h)-f(x,t)}{h}\,dx\right|\\ &+\left|\int_L^\infty D_tf(x,t)\,dx\right|.\end{aligned}$$

Next, we apply the fundamental theorem of calculus to $f(x,t+h)-f(x,t)$ to get

$$f(x,t+h)-f(x,t)=\int_t^{t+h} D_t f(x,s)\,ds,$$

so using the uniform integrability of $|D_t f(x,s)|$ we get, for $N>L$,

$$\begin{aligned}\left|\int_L^N \frac{f(x,t+h)-f(x,t)}{h}\,dx\right| &= \left|\int_L^N h^{-1}\int_t^{t+h} D_t f(x,s)\,dsdx\right| \\ &\le |h|^{-1}|\int_t^{t+h}\int_L^N D_t f(x,s)dxds| \\ &\le \epsilon/4.\end{aligned}$$

Therefore,

$$\left|\int_L^\infty \frac{f(x,t+h)-f(x,t)}{h}\,dx\right| = \left|\lim_{N\to\infty}\int_L^N \frac{f(x,t+h)-f(x,t)}{h}\,dx\right| \le \epsilon/4.$$

We also have

$$\left|\int_L^\infty D_t f(x,t)\,dx\right| < \epsilon/4.$$

For $\left|\int_a^L \left(\frac{f(x,t+h)-f(x,t)}{h} - D_t f(x,t)\right)dx\right|$, we apply our earlier argument to find a $\delta>0$ such that when $0<|h|<\delta$, we have

$$\left|\int_a^L \left(\frac{f(x,t+h)-f(x,t)}{h} - D_t f(x,t)\right)dx\right| < \epsilon/2.$$

This concludes our proof. ■

Example 5.7.3 Differentiation of a Poisson Integral.

Consider the Poisson integral

$$u(s,t)=\int_{\mathbb{R}} \frac{t\,u(x,0)}{(x-s)^2+t^2}dx$$

on the upper half plane $\{(s,t): t>0\}$, where $u(x,0)$ is a given bounded function.

To consider the differentiability of $u(s,t)$ for (s,t) near (s_0,t_0) with $t_0>0$, note that the integrand $f(x,s,t)=\frac{t\,u(x,0)}{(x-s)^2+t^2}$ is continuous for $(x,s,t)\in\mathbb{R}\times D_r(s_0,t_0)$, where $D_r(s_0,t_0)$ is the disk of radius r centered at (s_0,t_0) with $0<r<t_0/2$, so that when $(s,t)\in D_r(s_0,t_0)$, we have $t>t_0/2$; $f(x,s,t)$ is also differentiable in (s,t) for $(s,t)\in D_r(s_0,t_0)$.

To verify the uniform convergence of the improper integral $\int_{\mathbb{R}} f_s(x,s,t)\,dx$, we need to start with $\epsilon>0$, and look for $L>0$ such that for any $N>M>L$,

$$\left|\int_{N>|x|>M} f_s(x,s,t)\,dx\right| < \epsilon \quad \text{for all } (s,t)\in D_r(s_0,t_0).$$

In fact, to verify differentiability in s at (s_0,t_0), it suffices to check the above inequality for s near s_0 and $t=t_0$. The main focus will be to make sure that $f_s(x,s,t_0)$ tends to 0 at a sufficiently fast rate in x as $x\to\infty$, with uniform control on the coefficients for s near s_0, such that the above estimate holds; so we will watch out for how $f_s(x,s,t_0)$ depends on x and s.

Since

$$|f_s(x,s,t_0)| \le \frac{2t_0|s-x||u(x,0)|}{[(s-x)^2+t_0^2]^2} \le \frac{2t_0U|s-x|}{[(s-x)^2+t_0^2]^2},$$

where $U > |u(x,0)|$ for all $x \in \mathbb{R}$. We argue that when s is near s_0, and $|x| > M$, $|s-x|$ will be large; more precisely, for s is sufficiently near s_0, we have $|s| \le 2|s_0|$, and $|s-x| \ge |x| - |s| \ge M - 2|s_0| \ge M/2$, if M is chosen so that $M > 4|s_0|$, then

$$\begin{aligned}\left|\int_{N>|x|>M} f_s(x,s,t)\,dx\right| &\le \int_{|x-s|\ge M/2} \frac{2t_0U|s-x|}{[(s-x)^2+t_0^2]^2}dx \\ &\le \int_{M/2}^{\infty} \frac{4t_0Uz}{[z^2+t_0^2]^2}dz\end{aligned}$$

which can be made smaller than ϵ when M is sufficiently large, as the integral $\int_{M/2}^{\infty} \frac{z}{[z^2+t_0^2]^2}dz$ is convergent.

Note that t_0 is considered fixed in this argument; there will be difficulty to verify the uniform integrability if we allow $t \to 0$.

Exercise 5.7.4 Differentiation of the integral $F(t) = \int_{\mathbb{R}} \frac{dx}{(1+t^2x^2)(1+x^2)}$**.**

- Check whether $\int_{\mathbb{R}} D_t f(x,t)\,dx$ satisfies the uniform convergence criterion near $t > 0$ and $t = 0$, where $f(x,t)$ is the integrand.
- Determine whether $F'(0) = \int_{\mathbb{R}} D_t f(x,0)\,dx$.

5.8 Convexity and Some Applications

Convexity plays an important role in many extremal problems and inequalities. We include this section here to illustrate how some of the continuity and compactness arguments are used in the context of convex functions.

5.8.1 Convex sets vs Convex Functions

The notion of a convex set in $\mathbb{R}^n$ is more general than that of a convex function.

Definition 5.8.1 Convex set.

A set C in $\mathbb{R}^n$ is called convex, if for any $\mathbf{a}, \mathbf{b} \in C$ and any $t \in \mathbb{R}, 0 \le t \le 1$, we have $(1-t)\mathbf{a} + t\mathbf{b} \in C$.

Geometrically, the set $\{(1-t)\mathbf{a} + t\mathbf{b} : 0 \le t \le 1\}$ is the line segment in $\mathbb{R}^n$ with $\mathbf{a}, \mathbf{b}$ as its ends. A convex set needs not have any interior point.

Definition 5.8.2 Convex Function.

A real-valued function f defined on a convex set C is called convex, if for any $\mathbf{a}, \mathbf{b} \in C$ and any $t \in \mathbb{R}, 0 \le t \le 1$, we have

$$f((1-t)\mathbf{a} + t\mathbf{b}) \le (1-t)f(\mathbf{a}) + tf(\mathbf{b}).$$

f is called strictly convex if we have the strict inequality

$$f((1-t)\mathbf{a}+t\mathbf{b}) < (1-t)f(\mathbf{a})+tf(\mathbf{b}) \text{ for any } 0<t<1.$$

f is called concave if $-f$ is convex. Equivalently, the defining inequality above is reversed for a concave function.

Geometrically, if we construct a line in $\mathbb{R}^n \times \mathbb{R} \supset C \times \mathbb{R}$ through the points $(\mathbf{a}, f(\mathbf{a})), (\mathbf{b}, f(\mathbf{b}))$, then it has parametric equation

$$\mathbf{x} = ((1-t)\mathbf{a}+t\mathbf{b}, (1-t)f(\mathbf{a})+tf(\mathbf{b})),$$

so $(1-t)f(\mathbf{a})+tf(\mathbf{b})$ is the "height of the line above the point" $(1-t)\mathbf{a}+t\mathbf{b}$. When f is convex, $f((1-t)\mathbf{a}+t\mathbf{b})$ stays below the height at $(1-t)\mathbf{a}+t\mathbf{b}$ of the above line segment for $0 \leq t \leq 1$. Here[1] is a Desmos page illustrating this geometric property.

f is a convex function iff the set $\{(\mathbf{x}, y) : \mathbf{x} \in C, y \geq f(\mathbf{x})\}$ in $\mathbb{R}^n \times \mathbb{R}$, called the **epigraph** of f, is convex. Another characterization of a convex function is that for every real number c, the **sub level set** of f defined by $\{\mathbf{x} : f(\mathbf{x}) \leq c\}$ is a convex set.

Because of this relation, properties of convex functions can often be studied as properties of convex functions. We will later discuss briefly the notion of **supporting hyperplane** of a convex set and that of the graph of a convex function.

5.8.2 Review of Properties of Univariate Convex Functions

The illustration above in the previous subsection also includes a sketch of the argument that the slope of the secant lines on a convex function of a single variable is an increasing function. Geometrically it seems clear that if f is a convex function of a single variable on an interval that contains $a<b<c$, then $f(c)$ stays above the secant line through $(a, f(a), (b, f(b))$. This can be derived from the above property of secant lines: for $c>b$,

$$\frac{f(c)-f(a)}{c-a} \geq \frac{f(b)-f(a)}{b-a} \Leftrightarrow f(c) \geq f(a) + \frac{f(b)-f(a)}{b-a}(c-a).$$

These inequalities also hold when $c<a$.

The continuity of a convex function of one variable at an interior point is proved using these bounds by linear functions. Say, a is an interior point. Then there exist b, c in the domain such that $b>a>c$, and the above property of secant lines implies that for any $x, b>x>a$,

$$f(a) + \frac{f(c)-f(a)}{c-a}(x-a) \leq f(x) \leq f(a) + \frac{f(b)-f(a)}{b-a}(x-a).$$

Then the sandwich theorem implies that $f(x) \to f(a)$ as $x \to a+$. The direction when $x \to a-$ is done in a similar way.

Question 5.8.3 How can we extend this argument to higher dimensions?

A convex function is certainly continuous at an interior point when constrained along any one-dimensional lines, but there are infinitely many lines through any given point. Here we will see some ideas of compactness at play.

Solution. We will assume that the origin is in the interior of the domain of f; in fact, we will assume the domain of f includes the unit ball centered

[1] www.desmos.com/calculator/mbdd9ymzug

at the origin, and describe ideas to prove the continuity of f at the origin.

On key idea is that *there exist a finite number of points on the unit sphere*, in fact $2n$ points, such that any $\mathbf{x}$ with sufficiently small $\|\mathbf{x}\|$ can be written as

$$\mathbf{x} = s\bar{\mathbf{x}} \text{ for some } s, 0 \le s \le 1 \text{ and} \tag{5.8.1}$$
$$\bar{\mathbf{x}} = t_1\mathbf{a}_1 + \cdots + t_n\mathbf{a}_n \text{ for some } t_i \ge 0, t_1 + \cdots + t_n = 1, \tag{5.8.2}$$

where $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ are among the $2n$ points on the unit sphere. Technically

$$(x_1, x_2, \ldots, x_n) = |x_1|\tilde{\mathbf{e}}_1 + |x_2|\tilde{\mathbf{e}}_2 + \ldots + |x_n|\tilde{\mathbf{e}}_n,$$

where $\tilde{\mathbf{e}}_i = \operatorname{sgn}(x_i)\mathbf{e}_i$. If we set $s = |x_1| + |x_2| + \ldots + |x_n|$, and when $(x_1, x_2, \ldots, x_n) \ne (0, 0, \ldots, 0)$, $t_i = |x_i|/s$, and $\mathbf{a}_i = \tilde{\mathbf{e}}_i$, $\bar{\mathbf{x}} = t_1\tilde{\mathbf{e}}_1 + t_2\tilde{\mathbf{e}}_2 + \ldots + t_n\tilde{\mathbf{e}}_n$, then we get (5.8.1), (5.8.2). The condition $s \le 1$ is satisfied when $\|\mathbf{x}\| \le \frac{1}{\sqrt{n}}$, for, by the Cauchy-Schwarz inequality

$$s = |x_1| + \cdots + |x_n| \le \sqrt{n}\sqrt{x_1^2 + \cdots + x_n^2} \le 1.$$

It then follows that

$$\begin{aligned} f(\mathbf{x}) \le& [1-s]f(\mathbf{0}) + sf(\bar{\mathbf{x}}) \\ \le& [1-s]f(\mathbf{0}) + s\left[t_1 f(\mathbf{a}_1) + \cdots + t_n f(\mathbf{a}_n)\right] \\ \le& f(\mathbf{0}) + s\left[M - f(\mathbf{0})\right]. \end{aligned}$$

where M is chosen so that $f(\tilde{\mathbf{e}}_i) \le M$ for all i. This implies that

$$f(\mathbf{x}) - f(\mathbf{0}) \le (M - f(\mathbf{0}))(|x_1| + \cdots + |x_n|) \to 0 \text{ as } \mathbf{x} \to 0.$$

We can certainly cover $\mathbb{S}^{n-1}$ by a finite number of similarly constructed sets, and carry out this argument, which would allow us to prove that

$$\limsup_{\mathbf{x}\to\mathbf{0}} f(\mathbf{x}) \le f(0).$$

To prove $\liminf_{\mathbf{x}\to\mathbf{0}} f(\mathbf{x}) \ge f(0)$, we use the property of the secant lines as done in the one-dimensional case. For any $\mathbf{x}$ such that $\|\mathbf{x}\| \le \frac{1}{\sqrt{n}}$. We bound $f(\mathbf{x})$ from below by the secant line through $(-\frac{\mathbf{x}}{\|\mathbf{x}\|\sqrt{n}}, f(-\frac{\mathbf{x}}{\|\mathbf{x}\|\sqrt{n}})), (\mathbf{0}, f(\mathbf{0}))$:

$$f(\mathbf{x}) - f(\mathbf{0}) \ge \frac{f(\mathbf{0}) - f(-\frac{\mathbf{x}}{\|\mathbf{x}\|\sqrt{n}})}{(\sqrt{n})^{-1}}\|\mathbf{x}\|.$$

Since the slope $\frac{f(\mathbf{0}) - f(-\frac{\mathbf{x}}{\|\mathbf{x}\|\sqrt{n}})}{(\sqrt{n})^{-1}}$ has a lower bound due to the upper bound of $f(-\frac{\mathbf{x}}{\|\mathbf{x}\|\sqrt{n}})$, this allows us to conclude that $\liminf_{\mathbf{x}\to\mathbf{0}} f(\mathbf{x}) \ge f(0)$.

The secant line property of a convex function implies that, if $[a, a+\epsilon]$ is in the domain of a convex function, then the slope of the secant line $\frac{f(x)-f(a)}{x-a}$ has a limit as $x \to a+$, although this limit could be $-\infty$. If a is an interior point of the domain, then picking some $c < a$ in the domain implies a lower bound of $\frac{f(x)-f(a)}{x-a}$ in terms of $\frac{f(c)-f(a)}{c-a}$ when $x > a$, so in such a case, f has a finite **left derivative** $D_-f(a)$ and **right derivative** $D_+f(a)$ at a, and $D_-f(a) \le D_+f(a)$. Furthermore, for any

$k, D_-f(a) \le k \le D_+f(a)$,

$$\frac{f(x)-f(a)}{x-a} \ge D_+f(a) \ge k, \text{ for } x > a;$$
$$\frac{f(x)-f(a)}{x-a} \le D_-f(a) \le k, \text{ for } x < a.$$

This then implies that

$$f(x) \ge f(a) + k(x-a) \text{ for all } x \text{ in the domain of } f.$$

Since the right hand side, $f(a) + k(x-a)$, represents a straight line, the above inequality shows that *a convex function has a (linear)* ***support function*** *at any interior point of its domain.*

The support function property of a convex function can be used to give a simple proof of Jensen's inequality.

Theorem 5.8.4 Jensen's inequality.

Suppose that $f : X \mapsto (A, B)$ and $p(x) \ge 0$ is a density function on X, namely, $\int_X p(x)\,dx = 1$. Suppose that $\phi : (A, B) \mapsto \mathbb{R}$ is convex, then

$$\phi\left(\int_X f(x)p(x)\,dx\right) \le \int_X \phi(f(x))p(x)\,dx.$$

In words, "ϕ evaluated at the average of f is not more than the average of $\phi \circ f$."

Proof. Set $\bar{f} = \int_X f(x)p(x)\,dx$. It is easy to rule out the possibility that $\bar{f} = A$ or B, so we may assume that $A < \bar{f} < B$. Using the support property of ϕ at $\bar{f}$, there exists some k such that

$$\phi(y) \ge \phi(\bar{f}) + k(y - \bar{f}) \text{ for all } y \in (A, B).$$

Substituting y by $f(x)$, multiplying the above inequality by $p(x)$ and integrating over $x \in X$, we get

$$\int_X \phi(f(x))p(x)\,dx \ge \phi(\bar{f})\int_X p(x)\,dx + k\left(\int_X f(x)p(x)\,dx - \bar{f}\int_X p(x)\,dx\right).$$

The right hand side is simply $\phi(\bar{f})$, which proves the Jensen's inequality. ■

Commonly used cases of Jensen's inequalities include $\phi(y) = -\ln y$ or $y \ln y$. Proofs for Hölder's and Minkowski's inequalities also use convexity in crucial ways.

Exercise 5.8.5 Prove that $\ln\left(\int_X e^{u(x)}p(x)\,dx\right) \ge \int_X u(x)p(x)\,dx$ **for** $p(x) \ge 0, \int_X p(x)\,dx = 1$**.**

5.8.3 Some Properties of Convex Functions of Several Variables

When proving the continuity of a convex function of several variables, we already saw the complications for multi-dimensions. We do not intend to do a serious study of properties of convex functions of several variables, but only want to briefly discuss a few properties related to the notion of *supporting planes* to illustrate how the notion of compactness comes into play.

Definition 5.8.6 Supporting Hyperplane of a Convex Set.

Let K be a convex set, $\mathbf{x}_0$ be a point on the boundary of K. K is said to have a supporting hyperplane at $\mathbf{x}_0$ if there exists a vector $\mathbf{n}$ such that

$$(\mathbf{x}-\mathbf{x}_0)\cdot\mathbf{n}\geq 0 \text{ for all } \mathbf{x}\in K.$$

Definition 5.8.7 Supporting Hyperplane of a Convex Function.

Let K be a convex set and $f(\mathbf{x})$ be a convex function defined on K. Let $\mathbf{x}_0\in K$. The graph of f at $(\mathbf{x}_0, f(\mathbf{x}_0))$ is said to have a supporting hyperplane if there exists a non-zero vector $\mathbf{v}$ such that

$$f(\mathbf{x})\geq f(\mathbf{x}_0)+\mathbf{v}\cdot(\mathbf{x}-\mathbf{x}_0) \text{ for all } \mathbf{x}\in K.$$

Note that we use the same terminology in these two contexts, but they have a slight distinction, as illustrated by the simple example $f(x)=-\sqrt{x}$ on $[0,1]$. As a function it does not have a supporting hyperplane (a straight line here) at $x=0$, but its epigraph has a supporting hyperplane at $(0,0)$ (a vertical line).

For a convex function of a single variable we gave a proof of the existence of a supporting line at any interior point using the property of secant lines. We can apply this argument along any direction to a convex function of several variables, but it alone would not give us a supporting hyperplane at a point on the graph. The extension to multi-dimensions would necessarily involve some kind of limiting argument and compactness. We will discuss the following theorems.

Theorem 5.8.8 Existence of a Supporting Hyperplane of a Closed Convex Set.

Any boundary point of a closed convex set has a supporting hyperplane.

Proof. Let $\mathbf{x}_0\in K$ be a boundary point of the closed convex set K. Then there exists a sequence $\mathbf{x}_k\notin K, \mathbf{x}_k\to\mathbf{x}_0$. Each $\mathbf{x}_k$ also has a closest point $\mathbf{p}_k\in K$. This is done either by the Bolzano-Weierstrass compactness theorem or the parallelogram law of the Euclidean norm

$$2\|\frac{\mathbf{p}-\mathbf{q}}{2}\|^2=\|\mathbf{p}-\mathbf{x}_k\|^2+\|\mathbf{q}-\mathbf{x}_k\|^2-2\|\frac{\mathbf{p}+\mathbf{q}}{2}-\mathbf{x}_k\|^2$$

and the completeness of $\mathbb{R}^n$. This law shows that if $\mathbf{q}_l\in K$ is such that

$$\|\mathbf{q}_l-\mathbf{x}_k\|\to\inf\{\|\mathbf{x}-\mathbf{x}_k\| : \mathbf{x}\in K\},$$

then $\mathbf{q}_l$ is a Cauchy sequence, therefore has a limit, and that limit is in K.

Next we claim that

$$(\mathbf{p}_k-\mathbf{x}_k)\cdot(\mathbf{x}-\mathbf{p}_k)\geq 0 \text{ for all } \mathbf{x}\in K. \tag{5.8.3}$$

This follows from considering

$$h(t):=(t\mathbf{x}+(1-t)\mathbf{p}_k-\mathbf{x}_k)\cdot(t\mathbf{x}+(1-t)\mathbf{p}_k-\mathbf{x}_k).$$

Note that $h(t)=\|t\mathbf{x}+(1-t)\mathbf{p}_k-\mathbf{x}_k\|^2$, and $t\mathbf{x}+(1-t)\mathbf{p}_k\in K$ for $0\leq t\leq 1$, so $h(0)\leq h(t)$ for all $0\leq t\leq 1$. It follows that

$$h'(0)=2(\mathbf{p}_k-\mathbf{x}_k)\cdot(\mathbf{x}-\mathbf{p}_k)\geq 0.$$

Define $\mathbf{n}_k = (\mathbf{p}_k - \mathbf{x}_k)/\|\mathbf{p}_k - \mathbf{x}_k\|$. Then $\mathbf{n}_k$ is a sequence of unit vectors, so there exists a subsequence, still denoted by itself, and a limiting unit vector $\mathbf{n}$ such that $\mathbf{n}_k \to \mathbf{n}$. We also know that $\mathbf{p}_k \to \mathbf{x}_0$ as

$$\|\mathbf{p}_k - \mathbf{x}_0\| \le \|\mathbf{p}_k - \mathbf{x}_k\| + \|\mathbf{x}_k - \mathbf{x}_0\| \le 2\|\mathbf{x}_k - \mathbf{x}_0\|.$$

For each fixed $\mathbf{x} \in K$, dividing through both sides of (5.8.3) by $\|\mathbf{p}_k - \mathbf{x}_k\|$, and passing to the limit, we get

$$\mathbf{n} \cdot (\mathbf{x} - \mathbf{x}_0) \ge 0,$$

which is the inequality defining a supporting hyperplane.

In summary, the idea is that, in the absence of a direct construction of a supporting plane at the given point, one finds a relatively easy way to construct a supporting plane at a nearby, but unspecified point, and one then takes a limiting process to obtain a supporting plane at the given point. ∎

Theorem 5.8.9 Existence of a Supporting Hyperplane of the Graph of a Convex Function at an Interior Point.

Let $f(\mathbf{x})$ be a convex function defined on the convex set K. If $\mathbf{x}_0 \in K$ in an interior point of K, then the graph of f has a supporting hyperplane at $(\mathbf{x}_0, f(\mathbf{x}_0))$.

Proof. The epigraph $G_f = \{(\mathbf{x}, y) : \mathbf{x} \in K, y \ge f(\mathbf{x})\}$ is a convex set. Its closure $\overline{G_f}$ is a closed convex set, and $(\mathbf{x}_0, f(\mathbf{x}_0))$ is on the boundary of $\overline{G_f}$. By the previous theorem, there exists a non-zero vector $\mathbf{n} = (\mathbf{v}, c)$ such that

$$\mathbf{v} \cdot (\mathbf{x} - \mathbf{x}_0) + c(y - f(\mathbf{x}_0)) \ge 0 \text{ for all } (\mathbf{x}, y) \in \overline{G_f}.$$

Since $\mathbf{x}_0 \in K$ in an interior point of K, we claim that $c \ne 0$. For, otherwise, we would have

$$\mathbf{v} \cdot (\mathbf{x} - \mathbf{x}_0) \ge 0 \text{ for all } \mathbf{x} \in K,$$

which would force $\mathbf{v} = \mathbf{0}$.

Next we claim that $c > 0$. This is because $(\mathbf{x}_0, f(\mathbf{x}_0) + t) \in G_f$ for any $t > 0$, and the above inequality then forces $c > 0$. Now it follows that for any $\mathbf{x} \in K$, applying the above inequality for $y = f(\mathbf{x})$ implies that

$$f(\mathbf{x}) \ge f(\mathbf{x}_0) - c^{-1}\mathbf{v} \cdot (\mathbf{x} - \mathbf{x}_0),$$

which demonstrates a supporting hyperplane to the graph of f at $(\mathbf{x}_0, f(\mathbf{x}_0))$. ∎

It is possible to prove this theorem directly using the properties of a convex function, along the lines of proof for the one dimensional case. You should try to construct such a proof, at least for the two dimensional case.

Exercise 5.8.10 The tangent plane of a convex function at a differentiable point is a supporting plane to the graph of the function. Furthermore, if the point is in the interior of the domain, then it is the unique supporting plane.

Hint. If f denotes the function, and $Df(\mathbf{x}_0)$ denotes the gradient of f at $\mathbf{x}_0$, it may be geometrically easier to consider

$$g(\mathbf{x}) = f(\mathbf{x}) - f(\mathbf{x}_0) - Df(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0),$$

which is also convex.

We close this subsection by discussing a more subtle application of convex/

concave functions in an optimization problem.

Example 5.8.11

Using concavity to identify the minimum of $\frac{\left|\sum_{i=1}^{n} a_i b_i\right|}{\left(\sum_{i=1}^{n} a_i^2\right)^{1/2}\left(\sum_{i=1}^{n} b_i^2\right)^{1/2}}$ under some constraints on $(a_1, \ldots, a_n), (b_1, \ldots, b_n)$

The constraints will be $0 < a \le a_i \le A$, $0 < b \le b_i \le B$.

We introduce the new variables $u_i = a_i^2, v_i = b_i^2$, and reformulate the problem in terms of u_i, v_i. The quotient then becomes

$$\frac{\sum_i \sqrt{u_i v_i}}{\sqrt{(\sum_i u_i)(\sum_i v_i)}},$$

and the constraints become

$$a^2 \le u_i \le A^2; b^2 \le v_i \le A^2.$$

Our argument will be based on the following observation.

1. $\sqrt{uv}$ is a concave function in the first quadrant.
2. For any (u, v) in the rectangle $[a^2, A^2] \times [b^2, B^2]$, there exists unique (p, q) with $p, q \ge 0$, such that
$$(u, v) = p(a^2, B^2) + q(A^2, b^2).$$
3. In the set up above, we have
$$\sqrt{uv} \ge paB + qAb,$$
with equality iff (u, v) equals (a^2, B^2), or (A^2, b^2), equivalently, $(p, q) = (1, 0)$, or $(0, 1)$.

For the second item, note that for any (u, v) in the rectangle $[a^2, A^2] \times [b^2, B^2]$, there exists a unique $s > 0$, such that $s(u, v)$ lies on the diagonal from (a^2, B^2) to (A^2, b^2), which implies the existence of a unique $0 \le t \le 1$ such that

$$s(u, v) = (1 - t)(a^2, B^2) + t(A^2, b^2).$$

This then implies our desired relation.

We remark that in proving the last item above, only the (strict) concavity of $\sqrt{uv}$ along the diagonal from (a^2, B^2) to (A^2, b^2) is used. It is this last item that makes it possible to bound $\sum_i \sqrt{u_i v_i}$ from below.

Now for each (u_i, v_i), we find (p_i, q_i) according to the second item above

$$(u_i, v_i) = p_i(a^2, B^2) + q_i(A^2, b^2),$$

then we can bound the quotient as

$$\frac{\sum_i \sqrt{u_i v_i}}{\sqrt{(\sum_i u_i)(\sum_i u_i)}} \ge \frac{\sum_i (p_i aB + q_i Ab)}{\sqrt{[\sum_i (p_i a^2 + q_i A^2)][\sum_i (p_i B^2 + q_i b^2)]}}.$$

Setting $p = \sum_i p_i, q = \sum_i q_i$, $\alpha = A/a, \beta = B/b$, and after dividing both the numerator and denominator of the quotient on the right hand above by ab, it becomes

$$\frac{p\beta + q\alpha}{\sqrt{(p + q\alpha^2)(p\beta^2 + q)}},$$

and now the task is to find the infimum of this quotient when $p, q \ge 0$, and identify when equality can occur. This calculus problem can be solved in a routine way, and the answer is $\frac{2\sqrt{\alpha\beta}}{\alpha\beta+1}$.

5.9 Exercises

1. Suppose that A is an $m \times n$ matrix with rank n, and that $m \geq n$.

(a). Prove that there exists some $\lambda > 0$ such that $\|A\mathbf{x}\| \geq \lambda\|\mathbf{x}\|$ for all $\mathbf{x} \in \mathbb{R}^n$.

(b). Prove that there exists some $\delta > 0$ such that any $m \times n$ matrix B satisfying $\|B - A\| < \delta$ also has rank n.

2. Suppose that $\mathbf{f}$ is differentiable at $\mathbf{a} \in \mathbb{R}^n$ and takes values in $\mathbb{R}^m$ with $m \geq n$, and that $D\mathbf{f}(\mathbf{a})$ has rank n.

(a). Prove that there exist some $r > 0$ and $\lambda > 0$ such that $\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{a})\| \geq \lambda\|\mathbf{x} - \mathbf{a}\|$ for all $\mathbf{x} \in \mathbb{R}^n$ such that $\|\mathbf{x} - \mathbf{a}\| \leq r$.

(b). Prove that there exists some $\delta > 0$ such that for all $\mathbf{y} \in B(\mathbf{f}(\mathbf{a}), \delta)$, the open ball of radius δ centered at $\mathbf{f}(\mathbf{a})$,

$$\min_{\mathbf{x} \in \overline{B(\mathbf{a}, r)}} \|\mathbf{f}(\mathbf{x}) - \mathbf{y}\|$$

is attained at some $\mathbf{x}^* \in B(\mathbf{a}, r)$, and that

$$(\mathbf{f}(\mathbf{x}^*) - \mathbf{y}) \cdot (D\mathbf{f}(\mathbf{x}^*)\mathbf{h}) = 0 \quad \text{for all } \mathbf{h} \in \mathbb{R}^n.$$

$D\mathbf{f}(\mathbf{x}^*)\mathbf{h}$ is a tangent vector to the image of $\mathbf{f}$ at $\mathbf{f}(\mathbf{x}^*)$, so the above statement shows that $\mathbf{f}(\mathbf{x}^*)$ is closest to $\mathbf{y}$ among $\mathbf{f}(\mathbf{x}), \mathbf{x} \in \overline{B(\mathbf{a}, r)}$ and $\mathbf{f}(\mathbf{x}^*) - \mathbf{y}$ is orthogonal to all tangents to the image of $\mathbf{f}$ at $\mathbf{f}(\mathbf{x}^*)$.

(c). In the set up above, assume furthermore that $D\mathbf{f}(\mathbf{x})$ is continuous in $\mathbf{x}$, then prove that $r > 0$ can be adjusted so that $D\mathbf{f}(\mathbf{x})$ has rank n for all $\mathbf{x} \in B(\mathbf{a}, r)$. Assume in addition that $m = n$. Prove that $\mathbf{f}(\mathbf{x}) = \mathbf{y}$ in the above and $r > 0$ above can be adjusted so that such an $\mathbf{x}$ is unique in $B(\mathbf{a}, r)$.

3. Verify that the function $A \in \mathcal{M}^{n\times n} \mapsto A^2$ is a differentiable function, where $\mathcal{M}^{n\times n}$ is the space of $n \times n$ matrices. Then determine its derivative and check whether it is invertible in the case of $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ respectively.

Show that there exist neighborhoods U, V of $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ such that $S : U \mapsto V$ has an inverse which is differentiable, namely, any matrix in V has a uniquely defined square root in U. Can one do the same around $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$?

Hint. It may not be practical to work out the Jacobian matrix in the usual matrix form, but it suffices to obtain a linear function $L(H) \in \mathcal{M}^{n\times n}$ of $H \in \mathcal{M}^{n\times n}$ such that $\|(A + H)^2 - A^2 - L(H)\|/\|H\| \to 0$ as $\|H\| \to 0$.

4. Consider the Nth Fourier series partial sum $s_N(f; x)$ of $f \in C[-\pi, \pi]$ as a linear map from $C[-\pi, \pi]$ to itself. Using the integral expression for $s_N(f; x)$ to show that its operator norm is $\frac{1}{2\pi}\int_{-\pi}^{\pi} |D_N(t)|\, dt$.

5. Prove a lower bound of $\dfrac{\int_{x_1}^{x_2} f(x)g(x)\,dx}{\sqrt{\left(\int_{x_1}^{x_2} f(x)^2\,dx\right)\left(\int_{x_1}^{x_2} g(x)^2\,dx\right)}}$ when $f(x), g(x)$ are subject to positive upper and lower bounds. This problem is from #93 in Part II, Chapter 2 of Polya and Szegö's classic "Problems and Theorems in Analysis I"

Let a, A, b, B be positive numbers such that $a < A, b < B$. If the two functions $f(x)$ and $g(x)$ are integrable over the interval $[x_1, x_2]$, and $a \leq f(x) \leq A, b \leq g(x) \leq B$ on the interval. Then

$$\frac{\int_{x_1}^{x_2} f(x)g(x)\,dx}{\sqrt{\left(\int_{x_1}^{x_2} f(x)^2\,dx\right)\left(\int_{x_1}^{x_2} g(x)^2\,dx\right)}} \geq \frac{2}{\sqrt{\frac{AB}{ab}} + \sqrt{\frac{ab}{AB}}}.$$

Chapter 6

Integration in Several Variables

There are many aspects of integration in several variables. In this chapter we first discuss the definition and evaluation of the integral of well behaved *scalar functions* on simple domains in the Euclidean space $\mathbb{R}^n$ which are generalizations of rectangular boxes, triangles and tetrahedrons in two and three dimensions, called **hypercubes** and **simplexes** respectively. We then extend the discussion to **domains with piecewise C^1 boundary**. In the next chapter we will extend the discussion, where the domain of integration is a lower dimensional surface in the Euclidean space $\mathbb{R}^n$, and, first, the integrand is a scalar function as usual, then, the integrand is given in terms of *coupling* of a vector field and the tangent plane of the surface, or more formally, the integration of a differential form over a surface or a **singular chain**.

Below are some main issues to be addressed.

1. It is not too different from integration in one dimension if one only discusses integrals of a scalar function defined in a multi-dimensional rectangle. One can establish a similar integrability criterion and prove a version of the **Fubini Theorem**, which reduces the evaluation of the integral to iterated one variable integrals.

2. Complications come in when one tries to define integrals in more general domains in multi-dimensions. One main issue is how to properly define **partitions**. Partitions in the Riemann sense require that there is a well defined notion of area (volume) for the ***cells*** used in the partition, which is why traditionally we only confine to rectangular cells in partitions. Another aspect is that we want the cells to be *non-overlapping*, and this is relatively easy to achieve using rectangular cells.

   One way to resolve the issue of decomposing a general domain as a non-overlapping union of rectangular cells is to enclose the domain (required to be bounded in Riemann integration) by a rectangular box and extend the integrand to be 0 in the complement. This would require accounting for the contributions to the upper and lower integrals from those rectangular cells in the partition of the enclosing rectangular box that "*straddle*" between the domain and its complement. This is not too difficult to address, but one does need to control the **measure** of the boundary of the domain of integration.

3. In any integration theory one expects the integral $\int_A f$ to be linear in the integrand f, namely, for any integrable f, g, and constants a, b,

$$\int_A (af + bg) = a \int_A f + b \int_A g.$$

It is also natural to expect that when A can be decomposed as a non-overlapping union of two subsets $A_1 \cup A_2$,

$$\int_A f = \int_{A_1} f + \int_{A_2} f.$$

One also hopes that the converse holds: if the terms on the right hand side are well defined, then the left hand side should be well defined and equals the right hand side.

In one dimension this is certainly true when A_1, A_2 are intervals. But even in that context, Riemann integration theory places restrictions on these sets: if we take $A = [0,1], A_1 = \mathbb{Q} \cap [0,1], A_2 = [0,1] \setminus A_1$, then it is reasonable to argue that Dirichlet's non-Riemann integrable function on $[0,1]$ could be regarded as integrable on both A_1 and A_2 as it is a constant on either set, but its Riemann integral on $[0,1]$ is not defined. One key issue is that Riemann's integration theory does not allow general sets such as A_2 here as cells in the partition, and each partition has to be a *finite* partition. As a result one can't directly define Riemann integral on a domain with infinite length (area or volume) or when the integrand is unbounded.

4. A more serious issue is that *reasonable limits of Riemann integrable functions may not be Riemann integrable.* The rectification of this issue requires by Lebesgue's integration theory.

5. The simplistic idea of using "nice cells" to partition a domain of integration becomes more challenging when one tries to carry it out to surfaces or hypersurfaces. The issue of defining the area (volume) of cells on such surfaces has to be tackled first.

6. A basic difficulty in discussing the integration on a surface is how to properly define a surface. Unlike in the case of a curve, which can always be thought of as defined through a map on an interval, there is no simple or canonical domain which can be used to define a surface, except for a *patch* of a surface.

7. For the purpose of integration theory on a surface, it suffices that one can partition the surface into some non-overlapping union of patches such that each patch has a *parametric representation* through a map defined on a nice domain such as a rectangular box or simplex, and one defines the integration on each such patch on this nice domain through this parametrization.

8. The implementation of the above approach still requires substantial preparation work. First, the existence and construction of such a partition would require careful proof. Second, there is usually no canonical parametrization for a patch of a surface, so one has to verify that different parametrizations do not lead to different values of the integrals.

6.1 Basic Definitions and Properties for the Integral of a Scalar Function on a Multi-dimensional Rectangle

There is little difference between one and higher dimensions in the discussion of this and next section; differences will show when discussing evaluation of integrals and change of variables.

Definition 6.1.1

A set of the form $R = [a_1, b_1] \times \cdots \times [a_n, b_n]$, where $a_i < b_i$ for each i, is called a **hypercube** in $\mathbb{R}^n$. We also informally call it an n-dimensional rectangle (or rectangular box). We define its **volume** to be

$$|R| := (b_1 - a_1) \cdots (b_n - a_n).$$

When $a_i = b_i$ is allowed, and this equality holds for at least one i, we call such an R a degenerate hypercube (or informally a degenerate rectangle), and define its volume to be 0. A hypercube R has faces, edges and vertices obtained when one or more of the x_i variables is held as fixed at either a_i or b_i, which are degenerate hypercubes.

Suppose that $\mathcal{P}_i$ is a partition of $[a_i, b_i]$ for $1 \leq i \leq n$. Then these partitions $\mathcal{P}_i$'s form a partition $\mathcal{P}$ of $R = [a_1, b_1] \times \cdots \times [a_n, b_n]$ in the following sense:

(i). The cells of the partition $\mathcal{P}$ consist of

$$\{S := I_1 \times \cdots \times I_n : I_i \text{ is a subinterval of } \mathcal{P}_i\};$$

(ii). The union of the cells of $\mathcal{P} = R$;

(iii). The intersection of any two different cells of $\mathcal{P}$ is either the empty set, or a degenerate rectangle; and

(iv). $|R| = \sum_{S \in \mathcal{P}} |S|$.

We may denote this partition by $\mathcal{P}_1 \times \cdots \times \mathcal{P}_n$.

The **partition size** $\lambda(\mathcal{P})$ of $\mathcal{P} := \mathcal{P}_1 \times \cdots \times \mathcal{P}_n$ is defined as

$$\lambda(\mathcal{P}) := \max\{|I_i| : S := I_1 \times \cdots \times I_n \in \mathcal{P}\}.$$

Remark 6.1.2

Only the verification of the third and last property requires more than a few lines of argument. To prove (iii), note that a partition of R in our definition is defined in terms of a collection of partitions $\mathcal{P}_i$ for each constitutive factor. As a result, if two rectangles S_1, S_2 in a partition of R is non-empty, then for each i, the constitutive factor $I_{i,1}$ for S_1 and respectively $I_{i,2}$ for S_2 must have non-empty intersection. This means that $I_{i,1}$ and $I_{i,2}$ either share one common end point or are identical. We can now conclude that if two rectangles S_1, S_2 in a partition of R have non-empty intersection, then the intersection must be a degenerate hypercube.

For (iv), one easy way is to use an induction argument on n, the dimension. $R' := [a_2, b_2] \times \cdots \times [a_n, b_n]$ is an $(n-1)$-dimensional rectangle, and the partitions $\{\mathcal{P}_2, \ldots, \mathcal{P}_n\}$ form a partition $\mathcal{P}'$ of R'. Suppose that $\{I_{1,1}, \ldots, I_{1,k}\}$ are the subintervals of $\mathcal{P}_1$. In the summation $\sum_{S \in \mathcal{P}} |S|$, group S according to its constitutive factor in the first variable: each S has the form of $I_{1,j} \times S'$ for some $I_{1,j} \in \mathcal{P}_1$ and $S' \in \mathcal{P}'$, thus we have

$$\sum_{S \in \mathcal{P}} |S| = \sum_{I_{1,j} \in \mathcal{P}_1} \sum_{S' \in \mathcal{P}'} |I_{1,j} \times S'|.$$

By induction, we have

$$\sum_{S'\in\mathcal{P}'} |I_{1,j}\times S'| = |I_{1,j}| \sum_{S'\in\mathcal{P}'} |S'| = |I_{1,j}||R'|.$$

It now follows that

$$\sum_{S\in\mathcal{P}} |S| = \sum_{I_{1,j}\in\mathcal{P}_1} |I_{1,j}||R'| = (b_1-a_1)|R'| = |R|.$$

One could allow more general choices of rectangles in a partition; it would just make it harder to bookkeep such rectangles, and our definition suffices for our purpose of defining the integral of a function.

Definition 6.1.3

Suppose that f is a bounded real-valued function defined on the rectangle R, $\mathcal{P}$ a partition of R, and a point $\mathbf{x}_\alpha \in R_\alpha$ is chosen for each sub rectangle R_α of $\mathcal{P}$. We define the corresponding **Riemann sum** as

$$R(f,\mathcal{P},\{\mathbf{x}_\alpha\}) := \sum_\alpha f(\mathbf{x}_\alpha)|R_\alpha|.$$

Define

$$M_S(f) := \sup_{\mathbf{x}\in S} f(\mathbf{x}) \quad \text{and} \quad m_S(f) := \inf_{\mathbf{x}\in S} f(\mathbf{x})$$

as the supremum, and respectively infimum, of f on the rectangle S. Then the **upper sum** $U(f,\mathcal{P})$, and respectively the **lower sum** $L(f,\mathcal{P})$, of f on R with respect to the partition $\mathcal{P}$ is defined as

$$U(f,\mathcal{P}) := \sum_\alpha M_{S_\alpha}(f)|S_\alpha|, \quad L(f,\mathcal{P}) := \sum_\alpha m_{S_\alpha}(f)|S_\alpha|.$$

Note that

$$L(f,\mathcal{P}) \le R(f,\mathcal{P},\{\mathbf{x}_\alpha\}) \le U(f,\mathcal{P}).$$

We are interested in whether $R(f,\mathcal{P},\{\mathbf{x}_\alpha\})$ has a limit as $\lambda(\mathcal{P}) \to 0$ which is independent of how $\mathbf{x}_\alpha \in S_\alpha$ is chosen. It is easier to study whether $U(f,\mathcal{P})$, and respectively $L(f,\mathcal{P})$, has a limit as $\lambda(\mathcal{P}) \to 0$.

Definition 6.1.4

A partition $\mathcal{P}^* := \mathcal{P}_1^* \times \cdots \times \mathcal{P}_n^*$ is called a **refinement** of partition $\mathcal{P} := \mathcal{P}_1 \times \cdots \times \mathcal{P}_n$ if each $\mathcal{P}_i^*$ is a refinement of $\mathcal{P}_i$.

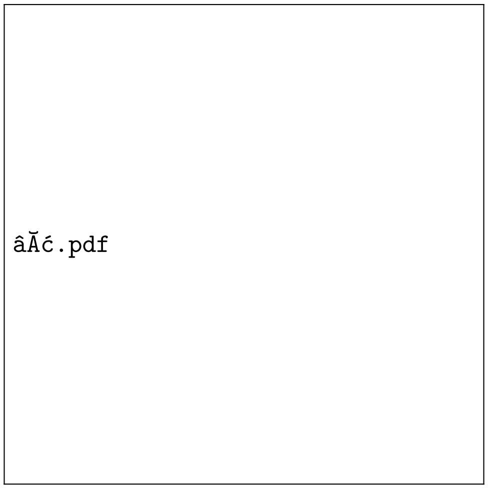


Figure 6.1.5 An illustration of a common refinement of two given partitions

Note that if $\mathcal{P}^*$ is a refinement of partition $\mathcal{P}$, then $\lambda(\mathcal{P}^*) \le \lambda(\mathcal{P})$, and that

$$U(f, \mathcal{P}^*) \le U(f, \mathcal{P}), L(f, \mathcal{P}^*) \ge L(f, \mathcal{P}).$$

Just as in one dimension, if $\mathcal{P}_1, \mathcal{P}_2$ are two partitions of R, then there is a partition $\mathcal{P}^*$ which is a refinement of both $\mathcal{P}_1$ and $\mathcal{P}_2$, and there is an essentially canonical way of constructing such an $\mathcal{P}^*$ by adjoining all the end points of the subintervals of the factors of $\mathcal{P}_1, \mathcal{P}_2$.

Definition 6.1.6

Suppose that f is a bounded function defined on the rectangle $R \subset \mathbb{R}^n$. Then its upper integral, and respectively lower integral, on R is defined as $\inf_{\mathcal{P}} U(f, \mathcal{P})$, and respectively $\sup_{\mathcal{P}} U(f, \mathcal{P})$, where $\mathcal{P}$ runs over all partitions of $\mathcal{P}$. The upper integral is denoted as $\overline{\int}_R f$, while the lower integral is denoted as $\underline{\int}_R f$.

Exercise 6.1.7 Let $\mathcal{C}$ denote the standard tertiary Cantor set on $[0, 1]$ and $\chi_{\mathcal{C}}$ denote its characteristic function which takes value 1 on $\mathcal{C}$ and 0 elsewhere. Let $\mathcal{P}$ be a partition of $[0, 1]$ into intervals of equal length 3^{-k} for some $k \in \mathbb{N}$. Find $U(\chi_{\mathcal{C}}, \mathcal{P})$ and $L(\chi_{\mathcal{C}}, \mathcal{P})$. Is there a positive lower bound of $L(\chi_{\mathcal{C}}, \mathcal{P})$ independent of $\mathcal{P}$?

Proposition 6.1.8 Basic Property of Upper and Lower Integral of a Bounded Real-valued Function Defined on a Rectangle.

Suppose that f is a bounded function defined on the rectangle $R \subset \mathbb{R}^n$. Then

$$\overline{\int}_R f = \lim_{\lambda(\mathcal{P}) \to 0} U(f, \mathcal{P}),$$

and

$$\underline{\int}_R f = \lim_{\lambda(\mathcal{P}) \to 0} L(f, \mathcal{P}).$$

Furthermore,

$$\underline{\int}_R f \le \overline{\int}_R f.$$

Proof. Let $\mathcal{P}_1, \mathcal{P}_2$ be two arbitrary partitions of R, and $\mathcal{P}^*$ be a refinement of both $\mathcal{P}_1$ and $\mathcal{P}_2$. Then

$$L(f, \mathcal{P}_1) \leq L(f, \mathcal{P}^*) \leq U(f, \mathcal{P}^*) \leq U(f, \mathcal{P}_2).$$

As a result,

$$L(f, \mathcal{P}_1) \leq \bar{\int}_R f = \inf_{\mathcal{P}_2} U(f, \mathcal{P}_2),$$

and

$$\underline{\int}_R f = \sup_{\mathcal{P}_1} L(f, \mathcal{P}_1) \leq \bar{\int}_R f.$$

For any $\epsilon > 0$, first find a partition $\mathcal{P}_2$ of R such that

$$\bar{\int}_R f \leq U(f, \mathcal{P}_2) \leq \bar{\int}_R f + \epsilon.$$

Now take any partition $\mathcal{P}$ of R of small partition size $\lambda(\mathcal{P})$ (to be determined). Let $\mathcal{P}^*$ be the canonical refinement of $\mathcal{P}$ and $\mathcal{P}_2$. Then

$$\bar{\int}_R f \leq U(f, \mathcal{P}_*) \leq U(f, \mathcal{P}_2) \leq \bar{\int}_R f + \epsilon.$$

and there exists a number N depending only on the partition $\mathcal{P}_2$ such that in the sums $U(f, \mathcal{P})$ and $U(f, \mathcal{P}^*)$, all but at most N terms may differ, as in the construction of $\mathcal{P}^*$, only those rectangles in the partition of $\mathcal{P}$ whose projection into some x_i variable contains a partition point corresponding to the projection of $\mathcal{P}_2$ need to be refined, and the number of such sub intervals of the factors of $\mathcal{P}$ that need to be refined in relation to $\mathcal{P}_2$ has a bound that depends only on $\mathcal{P}_2$.

Note that if S is such a rectangle, then after refinement in relation to $\mathcal{P}_2$, S is the non-overlapping union $\cup_{k=1}^{l} S_k$, so $|S| = \sum_{k=1}^{l} |S_k|$, and

$$0 \leq M_S(f)|S| - \sum_{k=1}^{l} M_{S_k}(f)|S_k| \leq 2 \max_R |f||S|$$

with $|S|$ bounded above by $(\lambda(\mathcal{P}))^n$. Thus

$$|U(f, \mathcal{P}) - U(f, \mathcal{P}^*)| \leq 2N \max_R |f|(\lambda(\mathcal{P}))^n.$$

It now follows that there exists some $\delta > 0$ such that whenever $\lambda(\mathcal{P}) < \delta$ we have

$$|U(f, \mathcal{P}) - U(f, \mathcal{P}^*)| < \epsilon \text{ so } U(f, \mathcal{P}) < \bar{\int}_R f + 2\epsilon,$$

proving that $\lim_{\lambda(\mathcal{P}) \to 0} U(f, \mathcal{P}) = \bar{\int}_R f$.

The proof of $\sup_{\mathcal{P}} L(f, \mathcal{P}) = \lim_{\lambda(\mathcal{P}) \to 0} L(f, \mathcal{P})$ is done in a similar way. ■

Exercise 6.1.9 Suppose that f, g are bounded functions on the rectangle R. Show that

$$L(f, \mathcal{P}) + L(g, \mathcal{P}) \leq L(f + g, \mathcal{P}) \text{ and } U(f + g, \mathcal{P}) \leq U(f, \mathcal{P}) + U(g, \mathcal{P}).$$

Definition 6.1.10

A bounded real-valued function f defined on a rectangle R is called **Riemann integrable**, if there exists a real number S such that $\lim_{\lambda(P)\to 0} R(f,\mathcal{P},\{\mathbf{x}_\alpha\}) = S$ in the sense that for any $\epsilon > 0$, there exists some $\delta > 0$ that

$$|R(f,\mathcal{P},\{\mathbf{x}_\alpha\}) - S| < \epsilon$$

for any partition $\mathcal{P}$ and choice of $\mathbf{x}_\alpha \in R_\alpha \in \mathcal{P}$, as long as $\lambda(P) < \delta$.

When such an S exists, it is unique, and we denote it by $\int_R f$.

Theorem 6.1.11 Linearity of Integral.

Suppose that f_1, f_2 are Riemann integrable on R, and c_1, c_2 are constants, then $c_1 f_1 + c_2 f_2$ is also Riemann integrable on R, and

$$\int_R (c_1 f_1 + c_2 f_2) = c_1 \int_R f_1 + c_2 \int_R f_2. \tag{6.1.1}$$

Proof. Let $S_1 = \int_R f_1$ and $S_2 = \int_R f_2$. Then for any $\epsilon > 0$, there exists $\delta > 0$ such that for any partition $\mathcal{P}$ of R, whenever $\lambda(\mathcal{P}) < \delta$, we have

$$\begin{aligned} |R(f_1,\mathcal{P},\{\mathbf{x}_\alpha\}) - S_1| < \epsilon, \\ |R(f_2,\mathcal{P},\{\mathbf{x}_\alpha\}) - S_2| < \epsilon. \end{aligned}$$

Then

$$\begin{aligned} &|R(c_1 f_1 + c_2 f_2,\mathcal{P},\{\mathbf{x}_\alpha\}) - (c_1 S_1 + c_2 S_2)\,| \\ \le & |c_1||R(f_1,\mathcal{P},\{\mathbf{x}_\alpha\}) - S_1| + |c_2||R(f_2,\mathcal{P},\{\mathbf{x}_\alpha\}) - S_2|| \\ < & \,(|c_1| + |c_2|)\,\epsilon, \end{aligned}$$

which shows the integrability of $c_1 f_1 + c_2 f_2$ as well as (6.1.1). ■

Based on Proposition 6.1.8, a *necessary condition* that f be Riemann integrable on R is that

$$\underline{\int}_R f = \overline{\int}_R f. \tag{6.1.2}$$

This turns out to be also *sufficient.*

Theorem 6.1.12 Riemann Integrability Criterion.

A bounded real-valued function f defined on the rectangle R is Riemann integrable iff (6.1.2) holds.

Proof. We only need to prove the if part. Suppose (6.1.2) holds. Call the value on both sides $\int_R f$. Our proof in Proposition 6.1.8 essentially carries over to show that, for any $\epsilon > 0$, there exists some $\delta > 0$, such that for any partition $\mathcal{P}$ of R with $\lambda(\mathcal{P}) < \delta$, we have

$$\int_R f - \epsilon < L(f,\mathcal{P}) \le \int_R f \le U(f,\mathcal{P}) < \int_R f + \epsilon.$$

Then for any choice of $\mathbf{x}_\alpha \in R_\alpha \in \mathcal{P}$, we have

$$\int_R f - \epsilon < L(f, \mathcal{P}) \le R(f, \mathcal{P}, \{\mathbf{x}_\alpha\}) \le U(f, \mathcal{P}) < \int_R f + \epsilon,$$

which shows that f is Riemann integrable on R. ■

Example 6.1.13

Let

$$f(x) = \begin{cases} 1 & x \in [0,1] \text{ rational}, \\ 0 & x \in [0,1] \text{ irrational}; \end{cases}$$

$$g(x,y) = \begin{cases} 1 & x \in [0,1] \text{ rational}, \; y = 0 \\ 0 & \text{elsewhere in } [0,1] \times [0,1]; \end{cases}$$

and $h(x,y) = f(x)f(y)$.

f is discontinuous everywhere on $[0,1]$, g is continuous on $[0,1] \times (0,1]$, but is discontinuous on $[0,1] \times \{0\}$, while h is discontinuous everywhere on $[0,1] \times [0,1]$.

For any partition $\mathcal{P}_1 : 0 = s_0 < s_1 < \cdots < s_k = 1$ of $[0,1]$ in the x variable, $U(f, \mathcal{P}_1) = 1$ and $L(f, \mathcal{P}_1) = 0$. Thus $\underline{\int}_{[0,1]} f(x)\,dx = 0$, $\overline{\int}_{[0,1]} f(x)\,dx = 1$, and f is not Riemann integrable on $[0,1]$.

For any partition $\mathcal{P}_2 : 0 = t_0 < t_1 < \cdots < t_l = 1$ of $[0,1]$ in the y variable, $m(g, [s_{i-1}, s_i] \times [t_{j-1}, t_j]) = 0$ for all i, j, and

$$M(g, [s_{i-1}, s_i] \times [t_{j-1}, t_j]) = \begin{cases} 1 & j = 0, \\ 0 & j > 1. \end{cases}$$

Thus $L(g, \mathcal{P}_1 \times \mathcal{P}_2) = 0$, $U(g, \mathcal{P}_1 \times \mathcal{P}_2) = t_1$, and $\underline{\int}_{[0,1]\times[0,1]} g = 0$, $\overline{\int}_{[0,1]\times[0,1]} g = 0$, so we conclude that g is Riemann integrable on $[0,1] \times [0,1]$, and $\int_{[0,1]\times[0,1]} g = 0$.

For h, we have $m(h, [s_{i-1}, s_i] \times [t_{j-1}, t_j]) = 0, m(h, [s_{i-1}, s_i] \times [t_{j-1}, t_j]) = 1$ for all i, j, so $L(h, \mathcal{P}_1 \times \mathcal{P}_2) = 0$, $U(g, \mathcal{P}_1 \times \mathcal{P}_2) = 1$. As a result, $\underline{\int}_{[0,1]\times[0,1]} h = 0$, $\overline{\int}_{[0,1]\times[0,1]} h = 1$, so h is not Riemann integrable on $[0,1] \times [0,1]$.

Exercise 6.1.14 Let $0 < r_k < 1$ be such that $\sum_{k=1}^{\infty} r_k < \infty$. Define a Cantor set $\mathcal{K}$ on $[0,1]$ by removing the middle r_k portion of the remaining interval at stage k. Let $\chi_{\mathcal{K}}$ denote its characteristic function. Is $\chi_{\mathcal{K}}$ Riemann integrable on $[0,1]$?

6.2 Further Riemann Integrability Criteria

6.2.1 Riemann Integrability Criterion in terms of the Oscillation of the Integrand

To find a more easily checkable criterion for (6.1.2) , we first make the following definition.

Definition 6.2.1

Let f be defined on R. The oscillation of a function f over the set $S \subset R$ is defined to be

$$\sup_S f - \inf_S f = M(f, S) - m(f, S)$$

and is denoted as $\operatorname{osc}(f, S)$.

The oscillation of a function f at a point $\mathbf{x}$ is defined to be

$$\lim_{r \searrow 0} \operatorname{osc}(f, R \cap B(\mathbf{x}, r)),$$

and is denoted as $\operatorname{osc}(f)(\mathbf{x})$. Here $B(\mathbf{x}, r)$ is the open ball of radius r centered at $\mathbf{x}$.

Note that, in the definition of $\operatorname{osc}(f)(\mathbf{x})$, we could replace the open ball $B(\mathbf{x}, r)$ by closed ball or rectangles. Furthermore,

$$\lim_{r \searrow 0} \operatorname{osc}(f, R \cap B(\mathbf{x}, r)) = \lim_{s \searrow 0} \operatorname{osc}(f, R \cap R(\mathbf{x}, s)) = \lim_{s \searrow 0} \operatorname{osc}(f, R \cap \bar{R}(\mathbf{x}, s)),$$

where $R(\mathbf{x}, s)$ denotes the open rectangle centered at $\mathbf{x}$ with $2s$ as its side length. This follows from the relation

$$\operatorname{osc}(f, U) \le \operatorname{osc}(f, V) \text{ whenever } U \subset V,$$

and $B(\mathbf{x}, r) \subset R(\mathbf{x}, r) \subset B(\mathbf{x}, \sqrt{n}r)$.

Note also that f is continuous at $\mathbf{x}$ iff $\operatorname{osc}(f)(\mathbf{x}) = 0$.

Exercise 6.2.2 Is it true that for any $\mathbf{x} \in U \subset R$ there holds $\operatorname{osc}(f)(\mathbf{x}) \le \operatorname{osc}(f, U)$?

Exercise 6.2.3 Is the following statement true: *if $\mathbf{x}$ is in a rectangle S, and $osc(f)(\mathbf{x}) \ge \epsilon$, then $M_S(f) - m_S(f) \ge \epsilon$?*

Proposition 6.2.4 Upper Semi-continuity of $\operatorname{osc}(f)(\mathbf{x})$.

The function $osc(f)(\mathbf{x})$ is upper semi-continuous. As a consequence, for any real number a, the set $\{\mathbf{x} : osc(f)(\mathbf{x}) \ge a\}$ is closed.

Proof. For any real number a, if $\mathbf{x}_0 \in \{\mathbf{x} : \operatorname{osc}(f)(\mathbf{x}) < a\}$, then there exists some $r > 0$ such that $\operatorname{osc}(f, B(\mathbf{x}_0, r)) < a$. For any $\mathbf{x} \in B(\mathbf{x}_0, r)$, we observe that $\operatorname{osc}(f)(\mathbf{x}) \le \operatorname{osc}(f, B(\mathbf{x}_0, r)) < a$. Thus $B(\mathbf{x}_0, r) \subset \{\mathbf{x} : \operatorname{osc}(f)(\mathbf{x}) < a\}$, proving that the latter is open. ■

Note that

$$\{\mathbf{x} : f \text{ discontinuous at } \mathbf{x}\} = \cup_{k \in \mathbb{N}} \{\mathbf{y} : \operatorname{osc}(f)(\mathbf{y}) \ge \frac{1}{k}\}, \tag{6.2.1}$$

namely, $\{\mathbf{x} : f \text{ discontinuous at } \mathbf{x}\}$ is a countable union of of the closed sets $\{\mathbf{y} : \operatorname{osc}(f)(\mathbf{y}) \ge \frac{1}{k}\}$.

Theorem 6.2.5 Riemann Integrability Criterion in terms of the Oscillation of the Integrand.

A bounded real-valued function f defined on the rectangle R is Riemann *integrable iff*

$$\forall \epsilon > 0, \exists \text{ a partition } \mathcal{P} := \{R_\alpha\} \text{ such that } \sum_\alpha osc(f, R_\alpha)|R_\alpha| < \epsilon. \quad (6.2.2)$$

In particular, if f is continuous on the closed rectangle R, then f is Riemann integrable on R.

Proof. Suppose that f is Riemann integrable on R and that $\epsilon > 0$ is given. Our proof of Proposition 6.1.8 gives us a partition $\mathcal{P} := \{R_\alpha\}$ such that

$$\int_R f - \frac{\epsilon}{2} < L(f, \mathcal{P}) \le \int_R f \le U(f, \mathcal{P}) < \int_R f + \frac{\epsilon}{2},$$

which implies that

$$\sum_\alpha \operatorname{osc}(f, R_\alpha)|R_\alpha| = U(f, \mathcal{P}) - L(f, \mathcal{P}) < \epsilon.$$

Suppose that (6.2.2) holds. Then $U(f, \mathcal{P}) - L(f, \mathcal{P}) < \epsilon$, and

$$0 \le \overline{\int_R} f - \underline{\int_R} f \le U(f, \mathcal{P}) - L(f, \mathcal{P}) < \epsilon.$$

Since $\epsilon > 0$ is arbitrary, it follows that $\overline{\int_R} f - \underline{\int_R} f = 0$, and f is Riemann integrable on R.

Finally, suppose that f is continuous on the closed rectangle R, then it is uniformly continuous on R. For any given $\epsilon > 0$, there exists $\delta > 0$ such that for any partition $\mathcal{P} := \{R_\alpha\}$ of R with $\lambda(P) < \delta$, we have $\operatorname{osc}(f, R_\alpha) < \epsilon/|R|$ for all $R_\alpha \in \mathcal{P}$. It then follows that

$$\sum_\alpha \operatorname{osc}(f, R_\alpha)|R_\alpha| \le \epsilon,$$

proving the Riemann integrability of f on R. ■

Note that when f is continuous on the closed rectangle R, we achieve (6.2.2) by constructing a partition $\mathcal{P} := \{R_\alpha\}$ of R such that $\operatorname{osc}(f, R_\alpha) < \epsilon/|R|$ for all $R_\alpha \in \mathcal{P}$. We can also achieve (6.2.2) if we can construct a partition $\mathcal{P} := \{R_\alpha\}$ of R such that the sum of volume of those rectangles R_α, for which $\operatorname{osc}(f, R_\alpha) \ge \epsilon/(2|R|)$, can be made smaller than $\epsilon/(2\operatorname{osc}(f, R))$, as the factors (f, R_α) have a uniform upper bound $\operatorname{osc}(f, R)$ for a given bounded function f on R. We will exploit some properties of $\operatorname{osc}(f)(\mathbf{x})$ for this.

Exercises

1. **Integrability of a real valued function in terms of the integrability of its positive and negative parts.** Let $f^+(x) = \max(f(x), 0), f^-(x) = \max(-f(x), 0)$ denote the positive and negative parts of $f(x)$ respectively. If f is Riemann integrable on C, is it true that f^+ and f^- are also Riemann integrable on C? If both f^+ and f^- are Riemann integrable on C, is it true that f is Riemann integrable on C?

2. Suppose that f is Riemann integrable on R. Prove that $|f|$ is Riemann integrable on R. Is the converse true?

3. Define $f(x,y)$ on $[0,1]\times[0,1]$ by

$$f(x,y)=\begin{cases}\frac{1}{q} & x \text{ rational and } y=\frac{p}{q} \text{ with } p,q \text{ co-prime;}\\ 0 & \text{otherwise}\end{cases}.$$

Show that $f(x,y)$ is Riemann integrable on $[0,1]\times[0,1]$ and $\int_{[0,1]\times[0,1]} f(x,y) = 0$.

6.2.2 Sets of Content 0 and of Measure 0

Definition 6.2.6

A set $S\subset\mathbb{R}^n$ is said to have content 0, if for any $\epsilon>0$, there exists a *finite* cover $\{R_i\}$ of S by rectangles such that

$$\sum_i |R_i| < \epsilon.$$

A set $S\subset\mathbb{R}^n$ is said to have measure 0, if for any $\epsilon>0$, there exists an *at most countable* cover $\{R_i\}$ of S by rectangles such that

$$\sum_i |R_i| < \epsilon.$$

Exercise 6.2.7 Set of measure 0 using covers of open rectangles. Verify that in the definition of a set of measure 0, open rectangles can be used instead pf closed rectangles. Watch out for places later on where this modification is needed.

Proposition 6.2.8 Relation Between Sets of Content 0 and Sets of Measure 0.

1. *A set of content* 0 *must be bounded and a set of measure* 0*.*
2. *The closure of a set of content* 0 *must be a set of content* 0*.*
3. *A countable union of sets of measure* 0 *must be a set of measure* 0*.*
4. *A countable union of sets of content* 0 *must be a set of measure* 0*.*
5. *A* compact *set of measure* 0 *is also a set of content* 0*.*

Proof. The first and second properties are obvious. For the third one, suppose $S=\cup_i S_i$, where each S_i is a set of measure 0. For any $\epsilon>0$, there exists an at most countable cover $\{R_{i\,j}\}_{j=1}^{N_i}$ (N_i could be ∞) of S_i such that $\sum_{j=1}^{N_i}|R_{i\,j}|<\frac{\epsilon}{2^i}$. Then $\{R_{i\,j}: i\in\mathbb{N}, 1\le j\le N_i\}$ is an at most countable cover of S, and

$$\sum_i\sum_{j=1}^{N_i}|R_{i\,j}| < \sum_i \frac{\epsilon}{2^i} = \epsilon,$$

proving that S is a set of measure 0.

The fourth property is a direct consequence of the third property. For the last property, suppose S is a compact set of measure 0. For $\epsilon>0$ we can choose a cover of S by *open* rectangles $\{R_i\}$ such that $\sum_i|R_i|<\epsilon$. Using the compactness of S, we can then select a finite sub cover which satisfies the desired property. ∎

Remark 6.2.9

The rectangles in the finite cover in the definition of a set of content 0 *are allowed to have non-empty overlaps and are not necessarily the cells of a partition, but we have the following*

Observation: *If* S *has content* 0*, then for any* $\epsilon > 0$*, there exists a partition* $\mathcal{P}$ *of a rectangle containing* S *such that*

$$\sum_{R_\alpha \in \mathcal{P}, R \cap S \neq \emptyset} |R_\alpha| < \epsilon. \tag{6.2.3}$$

This can be seen by first finding an open cover $\{U_\beta\}$ *of the closure* $\bar{S}$ *of* S *such that* $\sum_\beta |U_\beta| < \epsilon$*. Then for any* $\mathbf{x} \in \bar{S}$*, there exists some* $r_{\mathbf{x}} > 0$ *such that the open hypercube* $Q(\mathbf{x}, 4r_{\mathbf{x}})$ *with side lengths* $4r_{\mathbf{x}}$ *and centered at* $\mathbf{x}$ *is contained in* $\cup_\beta U_\beta$*. Considering the open cover* $\{Q(\mathbf{x}, r_{\mathbf{x}}) : \mathbf{x} \in \bar{S}\}$ *of* $\bar{S}$ *and using the compactness of* $\bar{S}$*, we find a finite cover* $\{Q(\mathbf{x}_i, r_{\mathbf{x}_i}) : 1 \le i \le N\}$*. Let* $\delta = \min\{r_{\mathbf{x}_i} : 1 \le i \le N\}$*. Then* $\delta > 0$ *and for any partition* $\mathcal{P}$ *of a rectangle* R *containing* $\bar{S}$ *such that* $\lambda(\mathcal{P}) < \delta$*, if any rectangle* R_α *of* $\mathcal{P}$ *satisfies* $R_\alpha \cap \bar{S} \neq \emptyset$*, then taking any* $\mathbf{x} \in R_\alpha \cap \bar{S}$*, there exists some* $\mathbf{x}_i$ *such that* $\mathbf{x} \in Q(\mathbf{x}_i, r_{\mathbf{x}_i})$*. This then implies that any point* $\mathbf{y} \in R_\alpha$ *also lies in* $Q(\mathbf{x}_i, 4r_{\mathbf{x}_i}) \subset \cup_\beta U_\beta$*. Thus* $\{R_\alpha : R_\alpha \cap \bar{S} \neq \emptyset\}$ *is a finite number of hypercubes in the partition* $\mathcal{P}$ *that covers* $\bar{S}$*, and* $\cup_{R_\alpha \cap \bar{S} \neq \emptyset} R_\alpha \subset \cup_\beta U_\beta$*, therefore*

$$\sum_{R_\alpha \in \mathcal{P}, R \cap \bar{S} \neq \emptyset} |R_\alpha| < \epsilon.$$

(6.2.3) then follows from this.

For a set S *of content* 0*, as a consequence of (6.2.3), we have*

$$U(\chi_S, \mathcal{P}) = \sum_{R_\alpha \in \mathcal{P}, R_\alpha \cap S \neq \emptyset} |R_\alpha| < \epsilon. \tag{6.2.4}$$

This then implies that $\int_R \chi_S = 0$*.*

Remark 6.2.10

The effect of requiring a finite cover in defining a set of content 0 *vs a possibly countably infinite cover in defining a set of measure* 0 *can be seen through the following examples.*

The set $\mathbb{Q}$ *of rationals in* $\mathbb{R}$ *is a set of measure* 0*, but not a set of content* 0*; its closure,* $\mathbb{R}$*, is not a set of measure* 0*. The set* $\mathbb{Z}$ *of integers is a closed set of measure* 0*, but not a set of content* 0*.*

Exercises

1. Show that if a set has content 0, then its boundary also has content 0.

2. Give an example of a closed set of measure 0 which does not have content 0 and an example of a bounded set of measure 0 such that its boundary does not have measure 0.

3. Suppose that f is an increasing function on $\mathbb{R}$. Show that the set of points where f is discontinuous has measure 0.

4. Does the Cantor set in Exercise 6.1.14 have content 0?

6.2.3 Riemann Integrability in terms of the Set of Discontinuity of the Integrand

We are now ready to formulate the following theorem.

Theorem 6.2.11 Riemann Integrability in terms of the Set of Discontinuity.

A bounded function f on a bounded closed rectangle R is Riemann integrable on R iff its set of discontinuity has measure 0.

Proof. We will use (6.2.1) for the only if part.

Suppose that f is Riemann integrable on R. For each $k \in \mathbb{N}$, we will prove that $D_k := \{\mathbf{y} : \operatorname{osc}(f)(\mathbf{y}) \geq \frac{1}{k}\}$ is a set of content 0.

Given any $\epsilon > 0$. There exists a partition $\mathcal{P} = \{R_\alpha\}$ of R such that

$$\sum_\alpha \operatorname{osc}(f, R_\alpha)|R_\alpha| < \frac{\epsilon}{2k}.$$

The rectangles in $\mathcal{P}$ are divided into two subgroups: the subgroup $\mathcal{L}_k$ consisting those R_α such that $\operatorname{osc}(f, R_\alpha) \geq \frac{1}{2k}$, and the subgroup $\mathcal{S}_k$ consisting those R_α such that $\operatorname{osc}(f, R_\alpha) < \frac{1}{2k}$. Then it follows from

$$\frac{1}{2k} \sum_{R_\alpha \in \mathcal{L}_k} |R_\alpha| \leq \sum_{R_\alpha \in \mathcal{L}_k} \operatorname{osc}(f, R_\alpha)|R_\alpha| < \frac{\epsilon}{2k}$$

that

$$\sum_{R_\alpha \in \mathcal{L}_k} |R_\alpha| < \epsilon.$$

We now claim that

$$D_k \subset \cup_{R_\alpha \in \mathcal{L}_k} R_\alpha.$$

This will show that D_k is a set of content 0.

If the claim were not true, there would exist some $\mathbf{x} \in D_k \setminus \cup_{R_\alpha \in \mathcal{L}_k} R_\alpha$. Thus $\mathbf{x} \in \cup_{R_\alpha \in \mathcal{S}_k} R_\alpha$. Since the complement of $\cup_{R_\alpha \in \mathcal{L}_k} R_\alpha$ is open, there exists some ball $B(\mathbf{x}, r) \subset \cup_{R_\alpha \in \mathcal{S}_k} R_\alpha$. If $\mathbf{x} \in$ interior(R_α) for some $R_\alpha \in \mathcal{S}_k$, it would force $\frac{1}{k} \leq \operatorname{osc}(f)(\mathbf{x}) \leq \operatorname{osc}(f, R_\alpha) < \frac{1}{2k}$, which would be a contradiction. So $\mathbf{x}$ can only be on the boundary of one or more $R_\alpha \in \mathcal{S}_k$. We can choose $r > 0$ small enough such that any $\mathbf{y} \in B(\mathbf{x}, r)$ and $\mathbf{x}$ will be in one such common rectangle. Therefore, $|f(\mathbf{y}) - f(\mathbf{x})| < \frac{1}{2k}$. This would lead to $\operatorname{osc}(f)(\mathbf{x}) \leq \operatorname{osc}(f, B(\mathbf{x}, r)) < \frac{1}{k}$, contradicting $\mathbf{x} \in D_k$.

The proof for the if part will be added. ■

Exercises

1. **The product of two Riemann integrable functions is Riemann integrable.** Let f, g be two Riemann integrable functions on R. Prove that their product $f \cdot g$ is Riemann integrable on R.
2. Is the characteristic function of the Cantor set in Exercise 6.1.14 Riemann integrable? What about the characteristic function of the complement of this Cantor set? ---Note that this complement is an open set of $\mathbb{R}$.

6.3 Fubini's Theorem

An integral is rarely evaluated using the limit of Riemann sum. Fubini's Theorem gives a mechanism to evaluate the integral of a function defined on a rectangle using iterated integrals in one variable. When the integrand is a continuous function on a rectangle, both the statement and proof of the theorem is straightforward. For the general case of a Riemann integrable function on a rectangle, the formulation and proof require some modification.

Suppose that $R_1 \subset \mathbb{R}^n$ and $R_2 \subset \mathbb{R}^m$ are two rectangles in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, and $f(\mathbf{x}, \mathbf{y})$ is a Riemann integrable function on $R := R_1 \times R_2$. Then for any fixed $\mathbf{x} \in R_1$, we can consider the integrability of the function $\mathbf{y} \mapsto f(\mathbf{x}, \mathbf{y})$ for $\mathbf{y} \in R_2$. We can also reverse the role between $\mathbf{x}$ and $\mathbf{y}$. To indicate the difference of the integration with respect to the variables, we write

$$\int_R f(\mathbf{x}, \mathbf{y})\, dA \text{ or } \int_R f(\mathbf{x}, \mathbf{y})\, d\mathbf{x}\, d\mathbf{y} \text{ for } \int_R f(\mathbf{x}, \mathbf{y}),$$

and

$$\int_{R_2} f(\mathbf{x}, \mathbf{y})\, d\mathbf{y} \text{ for the integral of } \mathbf{y} \mapsto f(\mathbf{x}, \mathbf{y}) \text{ over } \mathbf{y} \in R_2$$

when the integral exists. Likewise the upper and lower integrals

$$\overline{\int}_{R_2} f(\mathbf{x}, \mathbf{y})\, d\mathbf{y} \text{ and } \underline{\int}_{R_2} f(\mathbf{x}, \mathbf{y})\, d\mathbf{y}$$

make natural sense.

Theorem 6.3.1 Fubini's Theorem for Continuous Functions.

Suppose that $R_1 \subset \mathbb{R}^n$ and $R_2 \subset \mathbb{R}^m$ are two rectangles in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, and $f(\mathbf{x}, \mathbf{y})$ is a continuous function on $R := R_1 \times R_2$. Then $\int_{R_2} f(\mathbf{x}, \mathbf{y})\, d\mathbf{y}$ is a continuous function of $\mathbf{x} \in R_1$, and

$$\int_R f(\mathbf{x}, \mathbf{y})\, dA = \int_{R_1} \left(\int_{R_2} f(\mathbf{x}, \mathbf{y})\, d\mathbf{y} \right) d\mathbf{x}. \tag{6.3.1}$$

Likewise, $\int_{R_1} f(\mathbf{x}, \mathbf{y})\, d\mathbf{x}$ is a continuous function of $\mathbf{y} \in R_2$, and

$$\int_R f(\mathbf{x}, \mathbf{y})\, dA = \int_{R_2} \left(\int_{R_1} f(\mathbf{x}, \mathbf{y})\, d\mathbf{x} \right) d\mathbf{y}. \tag{6.3.2}$$

Proof. The key property used here is the uniform continuity of f on R: both sides of (6.3.1) and (6.3.2) can be approximated by any Riemann sum with respect to a partition whose size is sufficiently small. More specifically, for any $\epsilon > 0$, there exists $\delta > 0$ such that for any rectangle S_1 of R_1 and S_2 of R_2 whose side length is less than δ, we have

$$\operatorname{osc}(f(\mathbf{x}, \cdot), S_2) < \epsilon \text{ for any } \mathbf{x} \in R_1,$$

$$\operatorname{osc}(f(\cdot, \mathbf{y}), S_1) < \epsilon \text{ for any } \mathbf{y} \in R_2,$$

and

$$\operatorname{osc}(f, S_1 \times S_2) < \epsilon.$$

It follows that for any partition $\mathcal{P}_1$ of R_1 and $\mathcal{P}_2$ of R_2 whenever $\lambda(\mathcal{P}_1), \lambda(\mathcal{P}_2) < \delta$,

$$U(f(\mathbf{x}, \mathbf{y}), \mathcal{P}_2) - L(f(\mathbf{x}, \mathbf{y}), \mathcal{P}_2) < \epsilon |R_2| \text{ for any } \mathbf{x} \in R_1,$$

$$U(f(\mathbf{x},\mathbf{y}),\mathcal{P}_1) - L(f(\mathbf{x},\mathbf{y}),\mathcal{P}_1) < \epsilon|R_1| \text{ for any } \mathbf{y} \in R_2,$$

and

$$U(f(\mathbf{x},\mathbf{y}),\mathcal{P}_1\times\mathcal{P}_2) - L(f(\mathbf{x},\mathbf{y}),\mathcal{P}_1\times\mathcal{P}_2) < \epsilon|R_1||R_2|.$$

Since $\int_{R_1\times R_2} f(\mathbf{x},\mathbf{y})\,dA$ is sandwiched between $U(f(\mathbf{x},\mathbf{y}),\mathcal{P}_1\times\mathcal{P}_2)$ and $L(f(\mathbf{x},\mathbf{y}),\mathcal{P}_1\times\mathcal{P}_2)$, it follows that

$$\left|\int_{R_1\times R_2} f(\mathbf{x},\mathbf{y})\,dA - U(f(\mathbf{x},\mathbf{y}),\mathcal{P}_1\times\mathcal{P}_2)\right| < \epsilon|R_1||R_2|.$$

Likewise, $\int_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y}$ is sandwiched between $U(f(\mathbf{x},\mathbf{y}),\mathcal{P}_2)$ and $L(f(\mathbf{x},\mathbf{y}),\mathcal{P}_2)$, it follows that

$$\left|\int_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y} - U(f(\mathbf{x},\mathbf{y}),\mathcal{P}_2)\right| < \epsilon|R_2| \text{ for any } \mathbf{x} \in R_1,$$

therefore for any sampling $\{\mathbf{x}_\alpha\}$ of points corresponding to $\mathcal{P}_1$ we have

$$\left|\sum_{S_\alpha\in\mathcal{P}_1}\left(\int_{R_2} f(\mathbf{x}_\alpha,\mathbf{y})\,d\mathbf{y}\right)|S_\alpha| - \sum_{S_\alpha\in\mathcal{P}_1} U(f(\mathbf{x}_\alpha,\mathbf{y}),\mathcal{P}_2)|S_\alpha|\right| < \epsilon|R_1||R_2|.$$

Observe that $\sum_{S_\alpha\in\mathcal{P}_1} U(f(\mathbf{x}_\alpha,\mathbf{y}),\mathcal{P}_2)|S_\alpha|$ is sandwiched between $U(f(\mathbf{x},\mathbf{y}),\mathcal{P}_1\times\mathcal{P}_2)$ and $L(f(\mathbf{x},\mathbf{y}),\mathcal{P}_1\times\mathcal{P}_2)$, so

$$\left|\sum_{S_\alpha\in\mathcal{P}_1} U(f(\mathbf{x}_\alpha,\mathbf{y}),\mathcal{P}_2)|S_\alpha| - \int_{R_1\times R_2} f(\mathbf{x},\mathbf{y})\,dA\right| < \epsilon|R_1||R_2|.$$

It then follows that

$$\left|\sum_{S_\alpha\in\mathcal{P}_1}\left(\int_{R_2} f(\mathbf{x}_\alpha,\mathbf{y})\,d\mathbf{y}\right)|S_\alpha| - \int_{R_1\times R_2} f(\mathbf{x},\mathbf{y})\,dA\right| < 2\epsilon|R_1||R_2|.$$

But $\sum_{S_\alpha\in\mathcal{P}_1}\left(\int_{R_2} f(\mathbf{x}_\alpha,\mathbf{y})\,d\mathbf{y}\right)|S_\alpha|$ is a Riemann sum of the integral $\int_{R_1}\left(\int_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y}\right)\,d\mathbf{x}$ only subject to $\lambda(\mathcal{P}_1) < \delta$, it follows that

$$\left|\int_{R_1}\left(\int_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y}\right)\,d\mathbf{x} - \int_{R_1\times R_2} f(\mathbf{x},\mathbf{y})\,dA\right| \le 2\epsilon|R_1||R_2|. \tag{6.3.3}$$

Since $\epsilon > 0$ is arbitrary, it follows from (6.3.3) that (6.3.1) holds. (6.3.2) can be proved in a similar way. ■

Exercise 6.3.2 Verify that $\sum_{S_\alpha\in\mathcal{P}_1} U(f(\mathbf{x}_\alpha,\mathbf{y}),\mathcal{P}_2)|S_\alpha|$ is sandwiched between $U(f(\mathbf{x},\mathbf{y}),\mathcal{P}_1\times\mathcal{P}_2)$ and $L(f(\mathbf{x},\mathbf{y}),\mathcal{P}_1\times\mathcal{P}_2)$.

Exercise 6.3.3 Verify (6.3.3). What's the reason by replacing $<$ by $\le$?

To obtain a similar theorem for a general Riemann integrable function f on $R_1\times R_2$, the main issue is that the integrability of f on $R_1\times R_2$ does not necessarily guarantee that the iterated integrals $\int_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y}$, and respectively $\int_{R_1} f(\mathbf{x},\mathbf{y})\,d\mathbf{x}$, are defined for every $\mathbf{x}\in R_1$, and respectively for every $\mathbf{y}\in R_2$. One can see this through the simple example on $[0,1]\times[0,1]$

$$f(x,y) = \begin{cases} 1-\frac{1}{q} & x=\frac{p}{q}, p, q \text{ co-prime, and } y\in\mathbb{Q}, \\ 1 & \text{elsewhere.}\end{cases}$$

This $f(x,y)$ fails to be Riemann integrable in $y\in[0,1]$ for every $x\in\mathbb{Q}$. Yet, both

$\overline{\int}_0^1 f(x,y)\,dy$ and $\underline{\int}_0^1 f(x,y)\,dy$ are well defined for every $x \in [0,1]$. It turns out that we can formulate and prove a Fubini theorem using either the upper or the lower integral as a replacement in the iterated integral.

Exercise 6.3.4 Verify that the function $f(x,y)$ above is Riemann integrable on $[0,1]\times[0,1]$. The evaluate $\overline{\int}_0^1 f(x,y)\,dy$ and $\underline{\int}_0^1 f(x,y)\,dy$.

Here is a typical application of the Fubini's Theorem.

Example 6.3.5

Evaluate $\int_A e^{-x^3} y\,dx\,dy$ over the triangular region $A = \{(x,y) : 0 \le x \le 1, 0 \le y \le x\}$.

First the domain of integration is not a rectangle. We can treat this integral as the integral of $e^{-x^3} y\chi_A(x,y)$ over the square $R = \{(x,y) : 0 \le x \le 1, 0 \le y \le 1\}$.

We can try to evaluate the integral $\int_R e^{-x^3} y\chi_A(x,y)$ via

$$\int_0^1 \left(\int_0^1 e^{-x^3} y\chi_A(x,y)\,dx \right) dy = \int_0^1 \left(\int_x^1 e^{-x^3} y\,dx \right) dy,$$

but the integral $\int_x^1 e^{-x^3} y\,dx$ is not so easy to evaluate.

We then try to the alternate way of iterated integrals

$$\begin{aligned} \int_0^1 \left(\int_0^1 e^{-x^3} y\chi_A(x,y)\,dy \right) dx &= \int_0^1 \left(\int_0^x e^{-x^3} y\,dy \right) dx \\ &= \int_0^1 \left(\frac{1}{2} x^2 e^{-x^3} \right) dx \\ &= \frac{1}{6}\left(1 - e^{-1}\right). \end{aligned}$$

Exercise 6.3.6 Evaluate the integral $\int_R \frac{xy^2}{1+x^2+y^2}$, where $R = \{(x,y) : -1 \le x, y \le 1\}$.

Theorem 6.3.7 Fubini's Theorem for Riemann Integrable Functions.

Suppose that $R_1 \subset \mathbb{R}^n$ *and* $R_2 \subset \mathbb{R}^m$ *are two rectangles in* $\mathbb{R}^n$ *and* $\mathbb{R}^m$ *respectively, and* $f(\mathbf{x},\mathbf{y})$ *is a Riemann integrable function on* $R := R_1 \times R_2$. *Then both* $\overline{\int}_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y}$ *and* $\underline{\int}_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y}$ *are Riemann integrable of* $\mathbf{x} \in R_1$, *and*

$$\int_R f(\mathbf{x},\mathbf{y})\,dA = \int_{R_1} \left(\overline{\int}_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y} \right) d\mathbf{x} = \int_{R_1} \left(\underline{\int}_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y} \right) d\mathbf{x}. \tag{6.3.4}$$

Likewise, both $\overline{\int}_{R_1} f(\mathbf{x},\mathbf{y})\,d\mathbf{x}$ *and* $\underline{\int}_{R_1} f(\mathbf{x},\mathbf{y})\,d\mathbf{x}$ *are Riemann integrable of* $\mathbf{y} \in R_2$, *and*

$$\int_R f(\mathbf{x},\mathbf{y})\,dA = \int_{R_2} \left(\overline{\int}_{R_1} f(\mathbf{x},\mathbf{y})\,d\mathbf{x} \right) d\mathbf{y} = \int_{R_2} \left(\underline{\int}_{R_1} f(\mathbf{x},\mathbf{y})\,d\mathbf{x} \right) d\mathbf{y}.$$

Proof. We will focus on the first equality in (6.3.4). It relies on the observation that the lower and upper sums of $\overline{\int}_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y}$ for any partition $\mathcal{P}_1$ of R_1 and $\mathcal{P}_2$ of

R_2 are sandwiched between the lower sum $L(f(\mathbf{x},\mathbf{y}),\mathcal{P}_1\times\mathcal{P}_2)$ and the upper sum $U(f(\mathbf{x},\mathbf{y}),\mathcal{P}_1\times\mathcal{P}_2)$, and these two sums approach $\int_R f(\mathbf{x},\mathbf{y})\,dA$ when the partition size goes to zero.

$$\begin{aligned} L(f(\mathbf{x},\mathbf{y}),\mathcal{P}_1\times\mathcal{P}_2) \le L(\overline{\int}_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y},\mathcal{P}_1) \le U(\overline{\int}_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y},\mathcal{P}_1)\\ \le U(f(\mathbf{x},\mathbf{y}),\mathcal{P}_1\times\mathcal{P}_2). \end{aligned}$$

This follows by noting that

$$\sum_{S_{2,\beta}\in\mathcal{P}_2} m_{S_{2,\beta}}(f(\mathbf{x},\mathbf{y}))|S_{2,\beta}| \le \underline{\int}_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y}$$

and

$$\overline{\int}_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y} \le \sum_{S_{2,\beta}\in\mathcal{P}_2} M_{S_{2,\beta}}(f(\mathbf{x},\mathbf{y}))|S_{2,\beta}|$$

so

$$\begin{aligned} L(f(\mathbf{x},\mathbf{y}),\mathcal{P}_1\times\mathcal{P}_2) &= \sum_{S_{1,\alpha}\in\mathcal{P}_1}\sum_{S_{2,\beta}\in\mathcal{P}_2} m_{S_{1,\alpha}\times S_{2,\beta}}(f(\mathbf{x},\mathbf{y}))|S_{1,\alpha}||S_{2,\beta}|\\ &\le \sum_{S_{1,\alpha}\in\mathcal{P}_1}\sum_{S_{2,\beta}\in\mathcal{P}_2} m_{S_{1,\alpha}}(m_{S_{2,\beta}}(f(\mathbf{x},\mathbf{y}))|S_{1,\alpha}||S_{2,\beta}|\\ &\le \sum_{S_{1,\alpha}\in\mathcal{P}_1} m_{S_{1,\alpha}}(\underline{\int}_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y})|S_{1,\alpha}|\\ &\le \sum_{S_{1,\alpha}\in\mathcal{P}_1} m_{S_{1,\alpha}}(\overline{\int}_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y})|S_{1,\alpha}|\\ &\le \sum_{S_{1,\alpha}\in\mathcal{P}_1} M_{S_{1,\alpha}}(\overline{\int}_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y})|S_{1,\alpha}|\\ &\le \sum_{S_{1,\alpha}\in\mathcal{P}_1}\sum_{S_{2,\beta}\in\mathcal{P}_2} M_{S_{1,\alpha}}(M_{S_{2,\beta}}(f(\mathbf{x},\mathbf{y}))|S_{1,\alpha}||S_{2,\beta}|\\ &\le U(f(\mathbf{x},\mathbf{y}),\mathcal{P}_1\times\mathcal{P}_2). \end{aligned}$$

Since

$$\sup_{\mathcal{P}_1\times\mathcal{P}_2} L(f(\mathbf{x},\mathbf{y}),\mathcal{P}_1\times\mathcal{P}_2) = \inf_{\mathcal{P}_1\times\mathcal{P}_2} U(f(\mathbf{x},\mathbf{y}),\mathcal{P}_1\times\mathcal{P}_2) = \int_R f(\mathbf{x},\mathbf{y})\,dA$$

it follows that

$$\sup_{\mathcal{P}_1} L(\overline{\int}_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y},\mathcal{P}_1) = \inf_{\mathcal{P}_1} U(\overline{\int}_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y},\mathcal{P}_1)$$

proving that $\overline{\int}_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y}$ is Riemann integrable over $\mathbf{x}\in R_1$ with

$$\int_{R_1}\left(\overline{\int}_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y}\right)d\mathbf{x} = \int_R f(\mathbf{x},\mathbf{y})\,dA.$$

The rest of the equalities can be proved in a similar way. ■

Corollary 6.3.8 A Riemann integrable function $f(\mathbf{x},\mathbf{y})$ of $(\mathbf{x},\mathbf{y})$ is Riemann integrable in $\mathbf{y}$ except perhaps for $\mathbf{x}$ on a set of measure 0.

In the setting of the above theorem, $\overline{\int}_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y} = \underline{\int}_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y}$ *except perhaps on a set of measure* 0 *of* R_1*.*

Proof. It follows from (6.3.4) that

$$\int_{R_1}\left(\overline{\int}_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y} - \underline{\int}_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y}\right)d\mathbf{x} = 0.$$

Since $\overline{\int}_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y} - \underline{\int}_{R_2} f(\mathbf{x},\mathbf{y})\,d\mathbf{y} \geq 0$ for every $\mathbf{x} \in R_1$, the conclusion follows from the following exercise. ■

Exercise 6.3.9 Suppose that $g(\mathbf{x}) \geq 0$ on R_1 and $\overline{\int}_{R_1} g(\mathbf{x}) = 0$. Then $g(\mathbf{x}) = 0$ except perhaps on a set of measure 0 of R_1.

Exercise 6.3.10 Prove the rest of the equalities in Theorem 6.3.7.

6.4 Integration on a Jordan Measurable Set

We extend the Riemann integration from a rectangular region to a more general region called Jordan measurable.

Definition 6.4.1

A bounded subset G of $\mathbb{R}^n$ is called Jordan measurable if its boundary ∂G has content 0

Remark 6.4.2

Note that the characteristic function $\chi_G(\mathbf{x})$ *of* G *is discontinuous precisely on* ∂G*, so if* ∂G *has content* 0*, then* $\chi_G(\mathbf{x})$ *is Riemann integrable. Conversely, if* G *is bounded and* $\chi_G(\mathbf{x})$ *is Riemann integrable, then* ∂G *has measure* 0*. But* ∂G *is now a bounded closed set, so it has content* 0*, making* G *Jordan measurable.*

But this definition makes certain open sets not Jordan measurable, while it seems natural to define the length of any open set of $\mathbb{R}$ *to be* $\sum_k |I_k|$*, where* $\cup_k I_k$ *is the canonical decomposition of an open set* $\mathcal{O}$ *in* $\mathbb{R}$ *as the non-overlapping union of* open intervals.

Definition 6.4.3

Suppose that G of $\mathbb{R}^n$ is a Jordan measurable set and $f(\mathbf{x})$ is a bounded function on G. We say that f is integrable on G if $\chi_G(\mathbf{x})f(\mathbf{x})$ is a Riemann integrable function (on a rectangular box containing G). In such a case, we call $\int \chi_G(\mathbf{x})f(\mathbf{x})$ the integral of f on G and denote it as $\int_G f$.

Remark 6.4.4

The above definition is legitimate due to the following property: if S_1, S_2 are two rectangles containing G and $\chi_G(\mathbf{x})f(\mathbf{x})$ is Riemann integrable in S_1, then it is Riemann integrable in S_2 and $\int_{S_1} \chi_G(\mathbf{x})f(\mathbf{x}) = \int_{S_2} \chi_G(\mathbf{x})f(\mathbf{x})$.

Proposition 6.4.5 Basic Properties of Jordan Measurable Sets and Riemann Integrable Functions.

Suppose that A and B are Jordan measurable sets in $\mathbb{R}^n$. Then

(a) $A \cup B$, $A \cap B$, and $A \setminus B$ are Jordan-measurable.

(b) If A is Jordan-measurable and has measure 0, then any bounded function f on A is Riemann integrable on A and $\int_A f = 0$.

(c) If f, g are Riemann integrable on A, then for any constants a, b, $af + bg$ is Riemann integrable on A and

$$\int_A (af + bg) = a \int_A f + b \int_A g.$$

(d) If A, B are Jordan-measurable and $A \cap B$ has measure 0, then

$$\int_{A \cup B} f = \int_A f + \int_B f.$$

Proof. (a) follows by using $\partial(A \cup B) \subset \partial A \cup \partial B$, $\partial(A \cap B) \subset \partial A \cap \partial B$, and a similar one for $\partial(A \setminus B)$.

When A is Jordan-measurable and has measure 0, its interior must be empty so $A \subset \partial A$. Since ∂A has content 0, it follows that $\bar{A} = \partial A$ has content 0. Then, as in proving (6.2.4) and (6.2.4), for any $\epsilon > 0$, $\bar{A}$ can be covered by a finite number of open rectangles $\{S_i\}$ such that $\sum_i |S_i| < \epsilon$, so there exists some $\delta > 0$ such that if any partition $\mathcal{P}$ satisfies $\lambda(\mathcal{P}) < \delta$, then any rectangle R_α of $\mathcal{P}$ satisfying $R_\alpha \cap \bar{A} \neq \emptyset$ must satisfy $R_\alpha \subset \cup_i S_i$. Therefore

$$\sum_{R_\alpha \cap \bar{A} \neq \emptyset} |R_\alpha| \leq \sum_i |S_i| < \epsilon.$$

It further follows that

$$-C\epsilon \leq -C \sum_{R_\alpha \cap \bar{A} \neq \emptyset} |R_\alpha| \leq L(\chi_A f, \mathcal{P}) \leq U(\chi_A f, \mathcal{P}) \leq C \sum_{R_\alpha \cap \bar{A} \neq \emptyset} |R_\alpha| < C\epsilon,$$

where $C > 0$ is such that $|f(\mathbf{x})| \leq C$ for all $\mathbf{x} \in A$. Since $\epsilon > 0$ in these inequalities, they show that $\chi_A f$ is Riemann integrable with $\int_A f = \int \chi_A f = 0$. ■

Remark 6.4.6

The most common domains of integration are built on the following kinds: A is a Jordan-measurable set in $\mathbb{R}^{n-1}$, and $f(\mathbf{x}) \leq g(\mathbf{x})$ for all $\mathbf{x} \in A$ are two continuous functions on A, then the graph region $G_{A;f,g} := \{(\mathbf{x}, y) : \mathbf{x} \in A, f(\mathbf{x}) \leq y \leq g(\mathbf{x})\}$ is Jordan-measurable in $\mathbb{R}^n$.

If $h(\mathbf{x}, y)$ is a continuous function defined on $G_{A;f,g}$, then Fubini's Theo-

rem applied to $h(\mathbf{x}, y)\chi_{G_{A;f,g}}(\mathbf{x}, y)$ would take the form of

$$\int_{G_{A;f,g}} h(\mathbf{x}, y) = \int_A \left(\int_{f(\mathbf{x})}^{g(\mathbf{x})} h(\mathbf{x}, y)\, dy \right) d\mathbf{x}.$$

6.5 Determinant

This section is a brief review of determinants from linear algebra in preparation for the change of variables formula of integrals. The determinant is a scalar valued function $A \mapsto \det A$ on the space of square matrices with the following properties:

(a) $\det I_n = 1$ for any $n \times n$ identity matrix I_n.

(b) $A \mapsto \det A$ is a linear function of each of the column vectors of A when the other columns are held fixed.

(c) If B is obtained from A by interchanging two columns, then

$$\det B = -\det A.$$

(d) If A has two equal columns, then $\det A = 0$.

(e) If B is obtained from A by adding a multiple of one column of A to a different column, then $\det B = \det A$.

(f) $\det(AB) = \det A \det B$ for any two $n \times n$ matrices A, B.

In fact properties (d)-(f) follow from (a)-(c).

One can either produce a formula for $\det A$ in terms of the entries of A and verify that it satisfies the above properties, or use these properties to prove such a value of the determinant is uniquely determined---this will give an algorithm for computing $\det A$ based on the above properties. It turns out that the above properties indeed determine the $\det A$ uniquely. Its formula can be found in any standard textbook on linear algebra, but is rarely used directly except for 2×2 matrices.

These properties are rooted in the geometric origin in the determinant function.

1. Properties (a)-(e) encapsulate the properties of *signed* area (volume) of parallelograms (parallelepipeds) with edges formed using the column vectors of the matrix: property (e) is a reflection of the geometric property that parallelograms with equal base and equal heights have equal areas-the above operation corresponds to fixing an $(n-1)$ dimensional base and sliding the edge not in the base in the direction of a base edge, thus resulting in a newly formed parallelogram (parallelepiped) with the same base but equal height.

   It may be appealing to have a formula to directly compute the geometric area (volume) of a parallelogram (parallelepiped). But such a formula would lose the linearity as in (b). We would rather keep (b), and this necessitates in allowing negative values in the determinant, so the geometric interpretation has to be signed area (volume), which is related to the *orientation* of the edges of the parallelogram (parallelepiped)-for a 2×2 matrix, it reflects whether the relation between the first and second columns is counter-clockwise or clockwise rotation.

   We can still use $|\det A|$ to represent the geometric area (volume) of a parallelogram (parallelepiped). It fails the linearity but still maintains the following three geometric properties.

(i) If a multiple of one column of A is added to another column to form B, then $|\det B| = |\det A|$.

(ii) If one column of A is a linear combination of the rest of the columns of A, then $|\det A| = 0$. The converse also holds.

(iii) If one column of A is multiplied by a scalar c to form a matrix B, then $|\det B| = |c||\det A|$.

An equivalent way of stating property (ii) is that $\det A = 0$ iff the system $A\mathbf{x} = \mathbf{0}$ has a non-zero solution.

2. Any linear map associated to a square matrix A maps a region D to its image $A(D)$. Although we have not discussed the notion of area (volume) of a general region, the notion is intuitively clear at least when D is the non-overlapping union of rectangles (the interior of the rectangles are not allowed to overlap, but the edges are allowed to overlap). Then the ratio of the areas (volumes) of $A(D)$ and D is independent of D, and equals $|\det A|$. In other words, $|\det A|$ is the magnifying factor of area (volume) for A as a map. The property $\det(BA) = \det B \det A$ is a reflection of this perspective.

The following is a Desmos graph illustration[1] of areas of parallelograms with adjacent edges [u, v] and [u, v+pu].

6.6 Change of Variables in Multi-dimensional Integral

The change of variables formula in multi-dimensional integral takes the following form:

$$\int_{T(E)} f(\mathbf{y})\,d\mathbf{y} = \int_E f(T(\mathbf{x}))|\det DT(\mathbf{x})|\,d\mathbf{x} \tag{6.6.1}$$

where the change of variables transformation T is continuously differentiable in an open domain U such that $E \subset U$, and further conditions on f, E, T will be spelled out later.

Although an appropriate formulation and proof for general domains of integration takes a lot of work, there is a fairly intuitive idea behind (6.6.1), and the heart of the matter is reflected in the case of $f = 1$, which gives the volume $v(T(E))$ of $T(E)$, defined as $\int_{T(E)} 1\,d\mathbf{y}$

$$v(T(E)) = \int_{T(E)} 1\,d\mathbf{y} = \int_E |\det DT(\mathbf{x})|\,d\mathbf{x}. \tag{6.6.2}$$

Here is a version of the change of variables formula.

Theorem 6.6.1

Suppose that U is an open set in $\mathbb{R}^n$ and

$$T \text{ is continuously differentiable, injective on } U \text{ and } |\det DT(\mathbf{x})| > 0 \text{ for all } \mathbf{x} \in U. \tag{6.6.3}$$

Suppose that E is any Jordan-measurable subset of U such that its closure is in U. Then for any function f which is integrable on $T(E)$, (6.6.1) holds.

[1] www.desmos.com/calculator/895gqiflo9

Remark 6.6.2

The formulation of the above theorem implicitly assumes that $T(E)$ is Jordan measurable. This, together with the following, are consequences of (6.6.3) using the Inverse Function Theorem.

(i) $T(U)$ is an open set and there is a well defined inverse of T on $T(U)$, which is continuously differentiable.

(ii) For any subset $E \subset U$, $\partial\left(T(E)\right) = T(\partial E)$.

(iii) T maps any set of U of measure zero to a set of measure zero.

(iv) T maps any Jordan-measurable set of U to a Jordan measurable set.

Recall that the volume of a set E is defined as $\int_E \chi_E$ when the latter is well defined; this turns out to be equivalent to (a). ∂E has measure zero, and (b). $v(E) = \sup\left\{\sum_{S_i \subset E} v(S_i)\right\}$, where S_i are sub rectangles of a partition. (b) is simply a restatement of $v(E) = \sup L(\chi_E, \mathcal{P})$, noting that $L(\chi_E, \mathcal{P}) = \sum_{S_i \subset E} v(S_i)$ for any partition $\mathcal{P}$.

Remark 6.6.3

In the one-variable case, the change of variables transformation $y = T(x)$ is not required to be a bijection, and we don't put in the absolute value sign around the Jabobian, as our conventional notation encodes the orientation: if $T : [a, b] \mapsto [c, d]$, then

$$\int_{T(a)}^{T(b)} f(y)\, dy = \int_a^b f(T(x))T'(x)\, dx,$$

where, in case T is a bijection and $T(a) > T(b)$, we have $\int_{T(a)}^{T(b)} f(y)\, dy = -\int_{T(b)}^{T(a)} f(y)\, dy = -\int_{T[a,b]} f(y)\, dy$, which is consistent with

$$\int_{T[a,b]} f(y)\, dy = \int_{[a,b]} f(T(x))|T'(x)|\, dx.$$

Proof. (of properties in Remark 6.6.2) We only provide some details for item (iii). We first construct a sequence of compact closed subsets U_i of U satisfying $U = \cup_i U_i$. In fact we can take each U_i to be a hypercube with compact closure $\bar{U}_i$ in U and further require that $\bar{U}_i$ is contained in another hypercube V_i with compact closure $\bar{V}_i \subset U$.

Let $E \subset U$ be a set of measure 0. Then $E_i := E \cap U_i \subset U$ is a set of measure 0 and it suffices to prove that each $T(E_i)$ is a set of measure 0.

Since DT is continuous on U and $\bar{V}_i \subset U$ is compact and convex, there exists a bound $C_i > 0$ such that $\|DT(\mathbf{x})\| \le C_i$ for all $\mathbf{x} \in \bar{V}_i$. This then implies that

$$\|T(\mathbf{x}) - T(\mathbf{y})\| \le C_i\|\mathbf{x} - \mathbf{y}\| \text{ for any } \mathbf{x}, \mathbf{y} \in \bar{V}_i.$$

As a consequence, any hypercube Q in $\bar{V}_i$ with all side lengths equal to l is mapped by T to $T(Q)$, which is contained in a hypercube Q' of side length no more than $\sqrt{n}C_i l$ and $|Q'| \le n^{n/2}C_i^n|Q|$.

For any $\epsilon > 0$, let $\{Q_\alpha\}$ be a finite or countable collection of hypercubes covering E_i such that $\sum_\alpha |Q_\alpha| < \epsilon$. We may assume that each $Q_\alpha \subset U_i$---if some of the original Q_α does not satisfy this, we can simply replace it by \$$Q_\alpha \cap U_i$.

Claim: We can replace this collection, if necessary, by a collection of hypercubes $\{Q^*_\beta \subset V_i\}$, such that each Q^*_β is a hypercube with all side lengths equal and $\sum_\beta |Q^*_\beta| < 2\epsilon$.

This is seen by working with each individual Q_α: suppose it is given by $[a_1, b_1] \times \cdots [a_n, b_n]$, then choose $\sigma > 0$ and $k \in \mathbb{N}$ such that

$$(1+\sigma)^n \leq 1+\epsilon; 2^{1-k} < (b_j - a_j)\sigma \text{ for all } 1 \leq j \leq n.$$

By partitioning each axis into intervals of length 2^{-k}, we find $\hat{a}_j < \hat{b}_j \in 2^{-k}\mathbb{Z}$ such that

$$\hat{a}_j \leq a_j < b_j \leq \hat{b}_j, \quad \hat{b}_j - \hat{a}_j \leq b_j - a_j + 2^{1-k} \leq (b_j - a_j)(1+\sigma).$$

The hypercube $\hat{Q}_\alpha = [\hat{a}_1, \hat{b}_1] \times \cdots [\hat{a}_n, \hat{b}_n]$ contains Q_α with

$$|\hat{Q}_\alpha| = \Pi_{j=1}^n (\hat{b}_j - \hat{a}_j) \leq |Q_\alpha|(1+\epsilon).$$

Finally we can make sure that each $\hat{Q}_\alpha \subset V_i$ by working with a larger k, if necessary, and that each $\hat{Q}_\alpha$ is the union of a finite number of hypercubes of side length 2^{-k} with non-overlapping interior, therefore $|\hat{Q}_\alpha|$ is the sum of the volumes of these hypercubes. Collecting these hypercubes we get an at most countable collection $\{Q^*_\beta \subset V_i\}$ such that

$$\sum_\beta |Q^*_\beta| = \sum_\alpha |\hat{Q}_\alpha| < (1+\epsilon)\epsilon < 2\epsilon,$$

if we have chosen $1 > \epsilon > 0$.

Back to our proof of item (iii) in Remark 6.6.2. $T(E_i)$ is contained in $\cup_\beta T(Q^*_\beta)$, and each $T(Q^*_\beta)$ is contained in a cube Q'_β whose volume is $\leq n^{n/2} C_i^n |Q^*_\beta|$, therefore we have $\sum_\beta |Q'_\beta| \leq 2n^{n/2} C_i^n \epsilon$. Since n, C_i^n are fixed here and $\epsilon > 0$ is arbitrary, this shows that $T(E_i)$ has measure 0. ■

Remark 6.6.4

In the above argument we choose to work with nested hypercubes $U_i \subset \bar{U}_i \subset V_i \subset \bar{V}_i \subset U$ only to give us room in $\bar{V}_i$ to modify the hypercubes Q_α in the cover of E_i so that each is the non-overlapping union of a finite number of hypercubes of equal side lengths. Similar ideas can be used to prove the following lemma, which could have been used to simplify the above proof.

Lemma 6.6.5

Any open set U of $\mathbb{R}^n$ is a countable union of non-overlapping hypercubes of equal side lengths.

Corollary 6.6.6

Suppose that E in $\mathbb{R}^n$ is a set of measure zero. For any $\epsilon > 0$, there exists a countable union of non-overlapping hypercubes of equal side lengths Q_j covering E such that $\sum_j |Q_j| < \epsilon$.

Proof. (of (6.6.2)) The central idea in proving (6.6.2) is that for any open subset $U' \subset U$ with a compact closure $\bar{U}' \subset U$ and $\epsilon > 0$, there exists a $\delta > 0$ such that for any sufficiently small cube $Q \subset U'$ in the sense that its side length are equal and no more than δ, we have

$$v(T(Q)) \leq (1+\Lambda\epsilon)^n \left|\det\left(DT(\bar{\mathbf{x}})\right)\right| v(Q),$$

where $\bar{\mathbf{x}}$ is any point in Q, but will be taken as the center of Q, and $\Lambda \geq 1$ depends on T, U' such that

$$\Lambda^{-1}\|\mathbf{u}\| \leq \|DT(\mathbf{x})\mathbf{u}\| \leq \Lambda\|\mathbf{u}\|, \forall \mathbf{x} \in U', \mathbf{u} \in \mathbb{R}^n.$$

This is seen by the linear Taylor approximation of $T(\mathbf{x})$:

$$T(\mathbf{x}) = T(\bar{\mathbf{x}}) + \left(DT(\bar{\mathbf{x}}) + C(\mathbf{x},\bar{\mathbf{x}})\right)(\mathbf{x}-\bar{\mathbf{x}}) \text{ for any } \mathbf{x},\bar{\mathbf{x}} \in U',$$

where $|C(\mathbf{x},\bar{\mathbf{x}})| \leq \epsilon$ as long as $\|\mathbf{x}-\bar{\mathbf{x}}\| \leq \delta$ for some $\delta > 0$ which depends on T, U' and $\epsilon > 0$. For any cube Q in U' with $\bar{\mathbf{x}}$ as center and side lengths equal and no more than δ,

$$\{T(\bar{\mathbf{x}}) + DT(\bar{\mathbf{x}})(\mathbf{x}-\bar{\mathbf{x}}) : \mathbf{x} \in Q\}$$

is a parallelepiped with volume $|\det\left(DT(\bar{\mathbf{x}})\right)|v(Q)$.

We will use the Taylor expansion above to estimate the volume of $T(Q)$. It's easier to decompose $T = T_1 \circ T_2$ on Q, where T_1 is the linear map given by the matrix $(DT(\bar{\mathbf{x}}))$, and $T_2 = T_1^{-1} \circ T$. Note that the Jacobian matrix of T_2 at $\bar{\mathbf{x}}$ equal to the identity. If we apply T_1^{-1} to the Taylor expansion above we get

$$T_2(\mathbf{x}) = T_1^{-1} \circ T(\bar{\mathbf{x}}) + \left(I + T_1^{-1} \circ C(\mathbf{x},\bar{\mathbf{x}})\right)(\mathbf{x}-\bar{\mathbf{x}}).$$

It follows that $T_2(Q)$ is contained in a hypercube Q' with $T_1^{-1} \circ T(\bar{\mathbf{x}})$ as center and side lengths equal to the side lengths of Q multiplied by $(1+\Lambda\epsilon)$. Furthermore, $T(Q) = T_1 \circ T_2(Q) \subset T_1(Q')$ and $v(T_1(Q')) = |\det DT(\bar{\mathbf{x}})||Q'|$. Thus

$$v(T(Q)) \leq \left|\det\left(DT(\bar{\mathbf{x}})\right)\right|(1+\Lambda\epsilon)^n v(Q).$$

The approach here follows that in the article by *J. Schwartz, The formula for change in variables in a multiple integral, Amer. Math. Monthly 61, (1954), 81–85.*

The above estimate also holds for hypercubes whose ratios of side lengths are within 1 ± ϵ, with a somewhat modified constant replacing $(1+\Lambda\epsilon)^n$ and that constant still approaches 1 as $\epsilon \to 0$ — we will keep using $(1+\Lambda\epsilon)^n$ in the estimate for such hypercubes.

For any closed rectangle $S \subset U$ and any $\epsilon > 0$, we can do fine enough partition $\mathcal{P} = \{Q_\alpha\}$ of S using hypercubes for which the above estimate holds for each Q_α. Here we didn't use Lemma 6.6.5 to produce sub rectangles Q_α of S with equal side lengths as we want to work with a finite number of sub rectangles in a partition.

It now follows that

$$v(T(S)) = \sum_\alpha v(T(Q_\alpha)) \leq (1+\Lambda\epsilon)^n \sum_\alpha \left|\det\left(DT(\bar{\mathbf{x}_\alpha})\right)\right| v(Q_\alpha).$$

The summation above is a Riemann sum for the integral of $|\det\left(DT(\mathbf{x})\right)|$ on S, so as the partition size tends to 0, we get

$$v(T(S)) \leq (1+\Lambda\epsilon)^n \int_S \left|\det\left(DT(\mathbf{x})\right)\right| \, d\mathbf{x}.$$

Since $\epsilon > 0$ is arbitrary, it follows that

$$v(T(S)) \leq \int_S \left|\det\left(DT(\mathbf{x})\right)\right| \, d\mathbf{x}.$$

And this argument works not only for rectangles, but for all Jordan-measurable set. In fact, for any Jordan-measurable set E whose closure is in U, and any *non-negative* function f, integrable on $T(E)$,

$$\int_{T(E)} f(\mathbf{y})\,d\mathbf{y} \le \int_E f(T(\mathbf{x}))|\det DT(\mathbf{x})|\,d\mathbf{x}. \tag{6.6.4}$$

Here are some more details for proving (6.6.4). In defining $\int_{T(E)} f(\mathbf{y})\,d\mathbf{y}$, we may work with partitions $\mathcal{P}$ in the y-space such that any of its subrectangle that has non-empty intersection with $T(E)$ must be contained in $T(U)$. Then

$$\begin{aligned} L(f\chi_{T(E)},\mathcal{P}) &= \sum_{S_j:S_j\cap T(E)^c\neq\emptyset} m_{S_j}(f\chi_{T(E)})v(S_j) + \sum_{S_j:S_j\subset T(E)} m_{S_j}(f\chi_{T(E)})v(S_j) \\ &\le \sum_{S_j:S_j\subset T(E)} \int_{T^{-1}(S_j)} f(T(\mathbf{x}))|\det DT(\mathbf{x})|\,d\mathbf{x} \\ &\le \int_E f(T(\mathbf{x}))|\det DT(\mathbf{x})|\,d\mathbf{x}, \end{aligned}$$

where we have used $m_{S_j}(f\chi_{T(E)}) = 0$ when $S_j\cap T(E)^c \neq \emptyset$, $v(S_j) \le \int_{T^{-1}(S_j)} |\det DT(\mathbf{x})|\,d\mathbf{x}$ and $0 \le m_{S_j}(f\chi_{T(E)}) \le f(\mathbf{x})$ for $\mathbf{x} \in T^{-1}(S_j)$, as well as $\cup_{S_j\subset T(E)}T^{-1}(S_j)$ forming a non-overlapping subset of E. It follows that

$$\int_{T(E)} f(\mathbf{y})\,d\mathbf{y} = \sup_{\mathcal{P}} L(f\chi_{T(E)},\mathcal{P}) \le \int_E f(T(\mathbf{x}))|\det DT(\mathbf{x})|\,d\mathbf{x}.$$

■

Proof. (of (6.6.1)) First, since the set of discontinuity of $f(T(\mathbf{x}))|\det DT(\mathbf{x})|$ is $T^{-1}(D)$, where D is the set of discontinuity of $f(\mathbf{y})$ in $T(E)$, and $T^{-1}(D)$ has measure 0 due to D having measure 0, it follows that $f(T(\mathbf{x}))|\det DT(\mathbf{x})|$ is integrable on E.

It suffices to prove (6.6.1) for the case with $f \ge 0$. For the general case, we split an integrable function as the difference of its positive and negative parts: $f = f^+ - f^-$.

We now can apply (6.6.4) with T^{-1} on $T(E) \mapsto E$ and $f(T(\mathbf{x}))|\det DT(\mathbf{x})|$ as the integrand to obtain

$$\begin{aligned} \int_E f(T(\mathbf{x}))|\det DT(\mathbf{x})|\,d\mathbf{x} &\le \int_{T(E)} f(\mathbf{y})|\det DT(T^{-1}(\mathbf{y}))||\det DT^{-1}(\mathbf{y})|\,d\mathbf{y} \\ &= \int_{T(E)} f(\mathbf{y})\,d\mathbf{y}, \end{aligned}$$

where we have used $|\det DT(T^{-1}(\mathbf{y}))||\det DT^{-1}(\mathbf{y})| = 1$. This establishes (6.6.1) for non-negative integrable functions. ■

Remark 6.6.7

In applications often we can't apply the change of variables formula directly as the assumptions may not be satisfied in the form as stated, and we need to apply some approximation procedure.

One of the most commonly used change of variables is that from rectangular coordinates to polar coordinates:

$$\begin{bmatrix} x \\ y \end{bmatrix} = T\begin{bmatrix} r \\ \theta \end{bmatrix} = \begin{bmatrix} r\cos\theta \\ r\sin\theta \end{bmatrix}.$$

The Jacobian of T is $J_T(r,\theta) = r$. If $U = \{(r,\theta) : 0 < r < R, 0 < \theta < 2\pi\}$, *then T fails to be injective on a portion of ∂U and $T(U)$ is not quite the open disc $D_R = \left\{(x, y) : x^2 + y^2 < R^2\right\}$.*

However, for any $\epsilon > 0$ small, our Theorem is applicable on

$$U_\epsilon = \{(r, \theta) : \epsilon < r < R - \epsilon, \epsilon < \theta < 2\pi - \epsilon\}.$$

Thus for any function $f(x, y)$ which is continuous over $\bar{D}_R$,

$$\int_{T(U_\epsilon)} f(x, y)dxdy = \int_{U_\epsilon} f(r\cos\theta, r\sin\theta)rdrd\theta.$$

Then using

$$\int_{D_R} f(x, y)dxdy = \lim_{\epsilon\to 0} \int_{T(U_\epsilon)} f(x, y)dxdy,$$

and

$$\begin{aligned}&\int_U f(r\cos\theta, r\sin\theta)rdrd\theta\\ &= \lim_{\epsilon\to 0} \int_{U_\epsilon} f(r\cos\theta, r\sin\theta)rdrd\theta,\end{aligned}$$

we obtain

$$\int_{D_R} f(x, y)dxdy = \int_U f(r\cos\theta, r\sin\theta)rdrd\theta.$$

Just as in the one variable case, when f is not necessarily continuous (or bounded), one can define improper integral. An examination of the limiting argument above shows that if $f(x, y)$ is known to be continuous away from the origin, and for some $C > 0$ and $1 > \delta > 0$, we have

$$|f(x, y)| \le Cr^{-1-\delta},$$

then the improper integral $\int_{D_R} f(x, y)\, dxdy$ is well defined and the change of variables formula above is still valid.

6.7 A Brief Discussion on the Integration of Functions on a Surface

Examples of surfaces are easy to visualize, but the definition for a general surface is not easy, as there is no easy way to use a *single parametric representation* defined on a simple domain in the Euclidean space to represent a general surface. On the other hand, a local *patch* of a surface can always be represented by a single parametric map defined on a simple domain such as square or disc in the Euclidean space. Our focus will be on the study of such a local patch. But different parametric representations may represent the same surface patch, so we need to understand how the analytic and geometric properties of such a patch depend on the parametric representations.

The definition for the area of a surface and integral on a surface is even harder, as a general surface has no simple surface cells that have a natural and simple formula for their areas as rectangles do in a plane, so there is no easy way to define an integral on a surface directly using partitions of the surface.

We will use a parametric representation to define the area of a surface patch and verify that it is independent of the parametric representation used.

We will briefly discuss the integration of functions on a surface patch reducing it to the integration on a parametric cell.

6.7.1 Definition of an m-dimensional Differentiable Surface in $\mathbb{R}^n$

It seems natural to define a two dimensional surface as the graph in some $\mathbb{R}^n$ $(n > 2)$ of a continuously differentiable function defined over a two dimensional region in the flat Euclidean plane $\mathbb{R}^2 \subset \mathbb{R}^n$. But a surface as simple as the round sphere can't be represented as the graph of a single differentiable function. This graph feature becomes possible if one only requires this property locally near each point and allows for possibly different coordinate planes depending on the point over which to view the surface patch near the point as a graph.

Definition 6.7.1 Definition of an m-dimensional Differentiable Surface in $\mathbb{R}^n$.

A subset S of $\mathbb{R}^n$ is called an m-dimensional differentiable surface $(m < n)$, also called an m-dimensional **differentiable manifold**, if for each point $p \in S$, there exists a neighborhood $B(p, r)$ $(r > 0)$ of p in $\mathbb{R}^n$ and a differentiable function G defined in a neighborhood W of the coordinate projection p' of p to some m-dimensional coordinate subspace $\mathbb{R}^m$, where, for simplicity of notation, we take $p = (p_1, \cdots, p_n)$, $p' = (p_1, \cdots, p_m, 0, \cdots, 0)$, and W is open considered as a subset of $\mathbb{R}^m$, such that

$$S \cap B(p, r) = \{(\mathbf{x}, G(\mathbf{x})) : \mathbf{x} \in W \subset \mathbb{R}^m\}.$$

The conceptual difficulty of the above definition is that the function $G(\mathbf{x})$ and its domain W are not necessarily known explicitly or easily found and are allowed to vary with the point chosen, although they are often guaranteed by Theorem 5.5.8.

Proposition 6.7.2

Suppose that $F : U \subset \mathbb{R}^n \mapsto \mathbb{R}^{n-m}$ $(n > m)$ is continuously differentiable, that $\mathbf{u}_0 \in U$, and that the Jacobian matrix

$$DF(\mathbf{u}) \text{ has rank } n - m \text{ at every } \mathbf{u} \in U \text{ such that } F(\mathbf{u}) = F(\mathbf{u}_0). \tag{6.7.1}$$

Then the set $\{\mathbf{u} \in U : F(\mathbf{u}) = F(\mathbf{u}_0)\}$ is an m-dimensional surface containing $\mathbf{u}_0$.

Proof. This is a direction application of Theorem 5.5.8. ■

But (6.7.1) may not be easy to verify in concrete situations. It seems easier to consider the image of a continuously differentiable map defined over an m-dimensional region in the flat Euclidean plane $\mathbb{R}^m$ as a prototype over which to define a general m-dimensional surface. But in order for such a parametric representation to possess the usual properties of an m-dimensional surface, we need to assume that at each point the Jacobian matrix has rank m; this will guarantee that the image of the map does not degenerate into a lower dimensional object and indeed looks like an m-dimensional surface. Below is a more precise statement.

Proposition 6.7.3

Suppose that $U \subset \mathbb{R}^m$ is open, that

$$F : \mathbf{u} \in U \mapsto (x_1, \ldots, x_n) = (f_1(\mathbf{u}), \ldots, f_n(\mathbf{u})) \in \mathbb{R}^n$$

is continuously differentiable for some $n > m$, and that the Jacobian matrix $DF(\mathbf{u}_0)$ has rank m for some $\mathbf{u}_0 \in U$. Then there is a neighborhood $V \subset U$ of $\mathbf{u}_0$ such that $F(V) \subset \mathbb{R}^n$ can be represented as a continuously differentiable graph of $n - m$ of its variables in terms of the remaining m of its variables over a certain m-dimensional domain W in the Euclidean plane $\mathbb{R}^m$.

Proof. This is done as follows by applying the Inverse Function Theorem. For simplicity of notation, we will write out the case of $m = 2$. Here one may assume that $\frac{\partial(f_1,f_2)}{\partial(u,v)} \neq 0$ at (u_0, v_0), then one applies the Inverse Function Theorem to $(u, v) \mapsto (x_1, x_2) = (f_1(u, v), f_2(u, v))$ to show that it has a continuously differentiable inverse H defined on some open set W of $\mathbb{R}^2$ containing $(f_1(u_0, v_0), f_2(u_0, v_0))$ and $H(W) = V \subset U$ is an open neighborhood of (u_0, v_0). Then

$$F \circ H(x_1, x_2) = (x_1, x_2, f_3 \circ H(x_1, x_2), \ldots, f_n \circ H(x_1, x_2))$$

defines the desired graph over W. ■

It is often not necessary to find this graph representation explicitly, but work directly with the *parametric representation* $\mathbf{u} \in U \mapsto F(\mathbf{u})$. For a geometric surface, one either imposes that F be injective and has a continuous inverse on $F(U)$, or restricts to a small enough neighborhood so that this property holds. We call this a **surface patch** when $U \subset \mathbb{R}^m$ with $m = 2$ and for a general m a **manifold patch**.

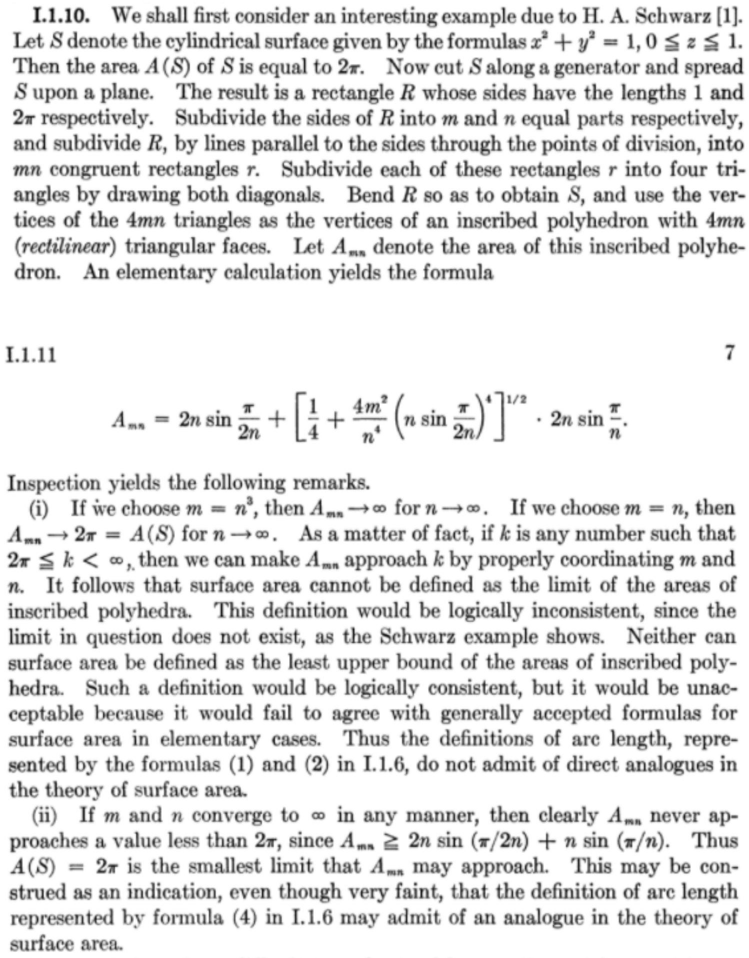
6.7.2 Volume of the Image of a Unit Hypercube under a Linear Map

Any notion of area of a surface or volume of a higher dimensional manifold must obey the universal property:

$$v(E \cup F) = v(E) + v(F) - v(E \cap F) \text{ when } v(E), v(F) \text{ are well defined.} \quad (6.7.2)$$

In addition, when two surfaces are obtained from each other by an *isometry* such as translation, rotation, reflection, or bending, they should have equal area. For example, a round cylinder of diameter 2 and height 1, when cut open along a generator on the side, unfolds into a rectangle of sides 1 and 2π, respectively, thus should have its area equal to 2π.

A naive generalization of the idea of defining the area of a surface as the least upper bound of the area of the inscribed triangulated surfaces turns out not valid. H. A. Schwarz found that even in the case of the round cylinder, the least upper bound of area of the inscribed triangulated surfaces is ∞. This is explained in Radó's *Length and Area*, AMS Colloquium Lectures, 1948.



I.1.10. We shall first consider an interesting example due to H. A. Schwarz [1]. Let S denote the cylindrical surface given by the formulas $x^2 + y^2 = 1, 0 \leqq z \leqq 1$. Then the area $A(S)$ of S is equal to 2π. Now cut S along a generator and spread S upon a plane. The result is a rectangle R whose sides have the lengths 1 and 2π respectively. Subdivide the sides of R into m and n equal parts respectively, and subdivide R, by lines parallel to the sides through the points of division, into mn congruent rectangles r. Subdivide each of these rectangles r into four triangles by drawing both diagonals. Bend R so as to obtain S, and use the vertices of the $4mn$ triangles as the vertices of an inscribed polyhedron with $4mn$ (*rectilinear*) triangular faces. Let A_{mn} denote the area of this inscribed polyhedron. An elementary calculation yields the formula

I.1.11 7

$$A_{mn} = 2n \sin\frac{\pi}{2n} + \left[\frac{1}{4} + \frac{4m^2}{n^4}\left(n \sin\frac{\pi}{2n}\right)^4\right]^{1/2} \cdot 2n \sin\frac{\pi}{n}.$$

Inspection yields the following remarks.

(i) If we choose $m = n^3$, then $A_{mn} \to \infty$ for $n \to \infty$. If we choose $m = n$, then $A_{mn} \to 2\pi = A(S)$ for $n \to \infty$. As a matter of fact, if k is any number such that $2\pi \leqq k < \infty$, then we can make A_{mn} approach k by properly coordinating m and n. It follows that surface area cannot be defined as the limit of the areas of inscribed polyhedra. This definition would be logically inconsistent, since the limit in question does not exist, as the Schwarz example shows. Neither can surface area be defined as the least upper bound of the areas of inscribed polyhedra. Such a definition would be logically consistent, but it would be unacceptable because it would fail to agree with generally accepted formulas for surface area in elementary cases. Thus the definitions of arc length, represented by the formulas (1) and (2) in I.1.6, do not admit of direct analogues in the theory of surface area.

(ii) If m and n converge to ∞ in any manner, then clearly A_{mn} never approaches a value less than 2π, since $A_{mn} \geqq 2n \sin(\pi/2n) + n \sin(\pi/n)$. Thus $A(S) = 2\pi$ is the smallest limit that A_{mn} may approach. This may be construed as an indication, even though very faint, that the definition of arc length represented by formula (4) in I.1.6 may admit of an analogue in the theory of surface area.

Figure 6.7.4 Radó's description of H. A. Schwarz's discovery that the least upper bound of area of the inscribed triangulated surfaces of a section of the round cylinder is infinity.

Let's next work out how to compute the area of a patch of a surface lying on an m-dimensional plane in $\mathbb{R}^n (n \geq m)$ as given by a parametric representation via a linear map. Let A be an $n \times m$ matrix with rank m and $m \leq n$. Consider

$$F : x \in \mathbb{R}^m \mapsto Ax \in \mathbb{R}^n.$$

$F(\mathbb{R}^m)$ is an $m-$dimensional subspace of $\mathbb{R}^n$. Let U be the standard cube in $\mathbb{R}^m$, then $F(U)$ is a parallelepiped in $F(\mathbb{R}^m)$, with $\text{col}_j(A)$, $j = 1, \ldots, m$ as edges. In the case of $m = 2$ and $n = 3$, $F(U)$ is a parallelogram with $\text{col}_1(A)$, $\text{col}_2(A)$ as edges and we know that its area can be computed as $|\text{col}_1(A) \times \text{col}_2(A)|$.

We now develop a formula for higher dimensional settings in the absence of cross product.

Theorem 6.7.5 Volume of the Image of a Unit Hypercube under a Linear Map.

In the setting above,

$$\textit{Volume}(F(U)) = \sqrt{\det(A^{\mathrm{T}}A)}\ \textit{Volume}(U). \tag{6.7.3}$$

Proof. Choose an orthonormal basis $\tau_1, \ldots, \tau_m$ of $F(\mathbb{R}^m)$, and express each $\operatorname{col}_j(A)$ in terms of this basis:

$$\operatorname{col}_j(A) = \sum_{i=1}^{m} B_{ij}\tau_i,$$

and let B be the $m \times m$ matrix (B_{ij}). Then

$$F(\mathbf{x}) = \sum_{i=1}^{m}\left(\sum_{j=1}^{m} B_{ij}x_j\right)\tau_i = [\tau_1 \cdots \tau_m]\, B\mathbf{x}$$

so in terms of the basis $\tau_1, \ldots, \tau_m$, $F(\mathbf{x})$ can be treated as given by $B\mathbf{x}$, and $F(U)$ can be thought of as obtained from the unit cube in $F(\mathbb{R}^m)$ by applying the linear transformation through multiplication by B, and by our earlier discussion

$$\text{Volume}(F(U)) = |\det B|\ \text{Volume}(U) = |\det B|.$$

On the other hand, the relations between A and B can be written as

$$A = [\tau_1 \cdots \tau_m]\, B,$$

so

$$A^{\mathrm{T}}A = B^{\mathrm{T}}\, [\tau_1 \cdots \tau_m]^{\mathrm{T}}\, [\tau_1 \cdots \tau_m]\, B = B^{\mathrm{T}}B,$$

using

$$[\tau_1 \cdots \tau_m]^{\mathrm{T}}\, [\tau_1 \cdots \tau_m] = I_m.$$

Now it follows that

$$|\det B| = \sqrt{\det B \det B^{\mathrm{T}}} = \sqrt{\det(B^{\mathrm{T}}B)} = \sqrt{\det(A^{\mathrm{T}}A)}$$

(B is square while A may not be square, so we can not have $\det(A^{\mathrm{T}}A) = \det(A^{\mathrm{T}})\det(A)$), and we arrive at (6.7.3). ■

Remark 6.7.6

Note that the (i,j) entry g_{ij} of $A^{\mathrm{T}}A$, is $\operatorname{col}_i(A) \cdot \operatorname{col}_j(A) = Ae_i \cdot Ae_j$. This is one of the geometric origins for the Riemannian metric tensor and the appearance of the area (volume) form $\sqrt{\det(g_{ij})}\, d\mathbf{u}$.

After we develop exterior algebra, we will find that

$$\mathit{col}_1(A) \wedge \mathit{col}_2(A) \wedge \cdots \wedge \mathit{col}_m(A) = (\det B)\ \tau_1 \wedge \tau_2 \wedge \cdots \wedge \tau_m,$$

so $\det B$ encodes the geometric relation between $\mathit{col}_1(A) \wedge \mathit{col}_2(A) \wedge \cdots \wedge \mathit{col}_m(A)$ and $\tau_1 \wedge \tau_2 \wedge \cdots \wedge \tau_m$ formed from two bases of $F(\mathbb{R}^m)$.

Example 6.7.7

The area of the parallelogram in $\mathbb{R}^n$ with $\mathbf{a} = (a_1, \cdots, a_n)$ and $\mathbf{b} = (b_1, \cdots, b_n)$ as its adjacent edges is given by

$$\sqrt{\det \begin{bmatrix} \mathbf{a}\cdot\mathbf{a} & \mathbf{a}\cdot\mathbf{b} \\ \mathbf{a}\cdot\mathbf{b} & \mathbf{b}\cdot\mathbf{b} \end{bmatrix}}.$$

Exercises

1. Find the area of the subset of the solution set of

$$\begin{cases} x_1 \quad - ax_3 - cx_4 &= 0 \\ \quad x_2 - bx_3 - dx_4 &= 0 \end{cases}$$

which orthogonally projects onto the unit square $\{(x_3, x_4) : 0 \le x_3, x_4 \le 1\}$ in the x_3–x_4 coordinate plane.

2. Find the areas of the triangles with $(1,0,0)$ and $(\cos(\frac{\pi}{n}), \pm\sin(\frac{\pi}{n}), \frac{1}{2m})$, and respectively with $(1,0,0)$ and $(\cos(\frac{\pi}{n}), \sin(\frac{\pi}{n}), \pm\frac{1}{2m})$, as vertices.

Answer. For the former it is $\sin(\frac{\pi}{2n})\sqrt{4^2\sin^4(\frac{\pi}{2n}) + \frac{1}{m^2}}$; for the latter it is $\frac{1}{m}\sin(\frac{\pi}{2n})$.

3. Find the three dimensional volume of the tetrahedron in $\mathbb{R}^4$ with $(1,0,0,0)$, $(0,1,0,0)$, $(0,0,1,0)$, $(0,0,0,1)$ as vertices.

Hint. The three dimensional volume of a tetrahedron equals 1/6 of the volume of a parallelepiped sharing three adjacent edges with the tetrahedron.

6.7.3 Volume of an m-dimensional Differentiable Surface in $\mathbb{R}^n$

Even if F is not given by a linear map, but is continuously differentiable, and its Jacobian matrix $DF(\mathbf{u})$ at any $\mathbf{u}$ has rank m, then we can still construct an orthonormal basis $\{\tau_1(\mathbf{u}), \cdots, \tau_m(\mathbf{u})\}$ of Span $\{D_1F(\mathbf{u}), \ldots, D_mF(\mathbf{u})\}$ (e.g. by the Gram-Schmidt orthogonalization procedure) and write

$$DF(\mathbf{u}) = [\tau_1(\mathbf{u}) \cdots \tau_m(\mathbf{u})]\, B(\mathbf{u})$$

where the entries of the $m \times m$ matrix $B(\mathbf{u})$ are continuous in $\mathbf{u}$; furthermore, the above discussion about the role of $g_{ij}(\mathbf{u}) = D_iF(\mathbf{u}) \cdot D_jF(\mathbf{u})$ still makes sense. Namely, $\sqrt{\det(g_{ij}(\mathbf{u}))}$ is the ratio of volume of $DF(\mathbf{u})(Q)$ and volume of Q, when Q is a hypercube in the parameter space $\mathbb{R}^m$ at $\mathbf{u}$, therefore, is the infinitesimal ratio of volume of $F(Q)$ and volume of Q when Q shrinks to u. This leads us to define

$$\int_U \sqrt{\det(g_{ij}(\mathbf{u}))}$$

as the **volume of** $F(U)$ for any (Jordan-measurable) open domain U of $\mathbb{R}^m$.

We can use the change of variables formula to check that this definition is independent of the parametrization, namely, if $\Phi : V \mapsto U$ is a diffeomorphism from V onto U, then $\hat{F}(\mathbf{v}) = F \circ \Phi(\mathbf{v}) : V \mapsto F(U)$ is another parametrization for $F(U)$, and we expect

$$\int_V \sqrt{\det(D\hat{F}(\mathbf{v})^{\mathrm{T}} D\hat{F}(\mathbf{v}))}\, d\mathbf{v} = \int_U \sqrt{\det(DF(\mathbf{u})^{\mathrm{T}} DF(\mathbf{u}))}\, d\mathbf{u}. \tag{6.7.4}$$

This follows by using the chain rule

$$D\hat{F}(\mathbf{v}) = DF(\Phi(\mathbf{v}))D\Phi(\mathbf{v}),$$

which leads to

$$D\hat{F}(\mathbf{v})^{\mathrm{T}} D\hat{F}(\mathbf{v}) = D\Phi(\mathbf{v})^{\mathrm{T}} DF(\Phi(\mathbf{v}))^{\mathrm{T}} DF(\Phi(\mathbf{v}))D\Phi(\mathbf{v}). \tag{6.7.5}$$

Since $D\Phi(\mathbf{v})^{\mathrm{T}}$, $DF(\Phi(\mathbf{v}))^{\mathrm{T}} DF(\Phi(v))$, and $D\Phi(v)$ are all $m \times m$ square matrices, so

$$\det(D\hat{F}(\mathbf{v})^{\mathrm{T}} D\hat{F}(\mathbf{v})) = \det D\Phi(\mathbf{v})^{\mathrm{T}} \det(DF(\Phi(\mathbf{v}))^{\mathrm{T}} DF(\Phi(\mathbf{v}))) \det D\Phi(\mathbf{v})$$

$$= \det(DF(\Phi(\mathbf{v}))^{\mathrm{T}} DF(\Phi(\mathbf{v})))|\det D\Phi(\mathbf{v})|^2.$$

This results in

$$\sqrt{\det(D\hat{F}(\mathbf{v})^{\mathrm{T}} D\hat{F}(\mathbf{v}))} = \sqrt{\det(DF(\Phi(\mathbf{v}))^{\mathrm{T}} DF(\Phi(\mathbf{v})))}|\det D\Phi(\mathbf{v})|, \qquad (6.7.6)$$

which, when applied to the left hand side of (6.7.4) in making the change of variables $\mathbf{u} = \Phi(\mathbf{v})$, confirms (6.7.4).

Remark 6.7.8

The computations above indicate that, if we accept the formal relation $d\mathbf{u} = |\det D\Phi(\mathbf{v})|d\mathbf{v}$ *in the sense of change of variables in integration, then*

$$\sqrt{\det(D\hat{F}(\mathbf{v})^{\mathrm{T}} D\hat{F}(\mathbf{v}))}\, d\mathbf{v} = \sqrt{\det(DF(\mathbf{u})^{\mathrm{T}} DF(\mathbf{u}))}\, d\mathbf{u}$$

is a geometric quantity independent of the parametrization, and is called the ***volume (area) form*** *of the submanifold (surface)* $F(U)$*, and denoted by* $dvol$ *or* dA*.*

Note also that if we define $\hat{g}_{ij}(\mathbf{v}) = D_i\hat{F}(\mathbf{v}) \cdot D_j\hat{F}(\mathbf{v})$*, then it is the* (i,j) *entry of* $D\hat{F}(\mathbf{v})^{\mathrm{T}} D\hat{F}(\mathbf{v})$*), and the relations above (6.7.5) and (6.7.6) become*

$$(\hat{g}_{ij}(\mathbf{v})) = D\Phi(\mathbf{v})^{\mathrm{T}}(g_{ij}(\Phi(\mathbf{v})))D\Phi(\mathbf{v}) \text{ and } \sqrt{\det(\hat{g}_{ij}(\mathbf{v}))}\, d\mathbf{v} = \sqrt{\det(g_{ij}(\mathbf{u}))}\, d\mathbf{u}.$$

The $\hat{g}_{ij}(\mathbf{v})$ *and* $g_{ij}(\mathbf{u})$ *here are obtained through* $D\hat{F}(\mathbf{v})$ *and* $DF(\mathbf{u})$ *respectively. There are other ways of obtaining such functions consisting of positive definite matrices associated with different parametrization and satisfying the above relations (6.7.5) and (6.7.6). Once we have them we can use them to define the volume of a region and the integral of a function just as above. The* $\hat{g}_{ij}(\mathbf{v})$ *and* $g_{ij}(\mathbf{u})$ *turn out to be the coordinate representations of a* ***Riemannian metric****.*

In the case of $m = 2$ and $n = 3$,

$$g_{11} = D_1F \cdot D_1F, g_{12} = D_1F \cdot D_2F, g_{22} = D_2F \cdot D_2F,$$

and we know from the properties of cross product and dot product for three dimensional vectors that

$$|D_1F \times D_2F| = \sqrt{\det(g_{ij})}$$

in this case.

Example 6.7.9

Suppose that U is a Jordan measurable open domain in $\mathbb{R}^n$, $g(\mathbf{x})$ has continuous and bounded partial derivatives in U. Then the graph of g over U has the representation $G(\mathbf{x}) = (\mathbf{x}, g(\mathbf{x}))$ with $D_iG(\mathbf{x}) = (\mathbf{e}_i, D_ig(\mathbf{x}))$, so $g_{ii} = 1 + |D_ig(\mathbf{x})|^2$ and $g_{ij} = D_ig(\mathbf{x})D_jg(\mathbf{x})$ for $i \neq j$. To find $\det(g_{ij}(\mathbf{x}))$, note that $(g_{ij}(\mathbf{x})) = I + [Dg(\mathbf{x})][Dg(\mathbf{x})]^t$, which has eigenvalues 1 with multiplicity $n-1$ and $1+|Dg(\mathbf{x})|^2$ with multiplicity 1, so $\det(g_{ij}(\mathbf{x})) = 1+|Dg(\mathbf{x})|^2$ and the volume of the graph of g over U is given by

$$\int_U \sqrt{1 + |Dg(\mathbf{x})|^2}\, d\mathbf{x}.$$

The requirement that $g(\mathbf{x})$ has continuous and bounded partial derivatives in U can be relaxed when the integral above can be treated as an appropriate improper integral, such as in the case of the area of a hemisphere when represented as a graph.

Example 6.7.10

The sphere $x^2+y^2+z^2=R^2$ in $\mathbb{R}^3$ is circumscribed by the round cylinder $x^2+y^2=R^2$. Both have parametric representation in terms of cylindrical coordinates, with the former given by

$$S(\theta,z)=(\sqrt{R^2-z^2}\cos\theta,\sqrt{R^2-z^2}\sin\theta,z),0\le\theta<2\pi,-R\le z\le R,$$

and the latter by

$$C(\theta,z)=(R\cos\theta,R\sin\theta,z),0\le\theta<2\pi,z\in\mathbb{R}.$$

Note that

$$D_\theta S(\theta,z)\cdot D_z S(\theta,z)=0,\|D_\theta S(\theta,z)\|=\sqrt{R^2-z^2},\|D_z S(\theta,z)\|=\frac{R}{\sqrt{R^2-z^2}},$$

so the area of the section of the sphere for $0\le\theta<2\pi,-R\le z_1\le z\le z_2\le R$ is given by

$$\int_0^{2\pi}\int_{z_1}^{z_2}\sqrt{\det\begin{bmatrix}D_\theta S(\theta,z)\cdot D_\theta S(\theta,z) & D_\theta S(\theta,z)\cdot D_z S(\theta,z)\\ D_\theta S(\theta,z)\cdot D_z S(\theta,z) & D_z S(\theta,z)\cdot D_z S(\theta,z)\end{bmatrix}}\,dz\,d\theta$$
$$=\int_0^{2\pi}\int_{z_1}^{z_2}R\;dz\,d\theta,$$

while

$$D_\theta C(\theta,z)\cdot D_z C(\theta,z)=0,\|D_\theta C(\theta,z)\|=R,\|D_z C(\theta,z)\|=1,$$

so the area of the section of the cylinder for $0\le\theta<2\pi,z_1\le z\le z_2$ is given by

$$\int_0^{2\pi}\int_{z_1}^{z_2}\sqrt{\det\begin{bmatrix}D_\theta C(\theta,z)\cdot D_\theta C(\theta,z) & D_\theta C(\theta,z)\cdot D_z C(\theta,z)\\ D_\theta C(\theta,z)\cdot D_z C(\theta,z) & D_z C(\theta,z)\cdot D_z C(\theta,z)\end{bmatrix}}\,dz\,d\theta$$
$$=\int_0^{2\pi}\int_{z_1}^{z_2}R\;dz\,d\theta.$$

The areas of these two sections are equal for any $-R\le z_1<z_2\le R$, which was first discovered by Archimedes. In particular, the area of the sphere is $\int_0^{2\pi}\int_{-R}^{R}R\;dz\,d\theta=4\pi R^2$.

One could represent the upper hemisphere as a graph $z=g(x,y)=\sqrt{R^2-x^2-y^2}$ and use the integration of $\sqrt{1+|Dz|^2}$ over $x^2+y^2<R$ to compute its area, but $|Dz|$ becomes unbounded as $x^2+y^2\to R^2$, so one has to deal with that issue. In the case here one needs to evaluate the improper integral $\int_{x^2+y^2<R^2}\frac{R}{\sqrt{R^2-x^2-y^2}}\,dxdy$, which can be converted into an improper integral in polar coordinate as $\int_0^{2\pi}\int_0^R\frac{Rr}{\sqrt{R^2-r^2}}\,drd\theta$.

Exercises

1. Find the area of the region of the plane $x + y + z = 1$ enclosed within the cylinder $x^2 + y^2 \leq 1$.

2. Find the area of the cylinder $x^2 + y^2 = 1$ intersected between the planes $x + y + z = \pm 1$.

3. Find the three dimensional volume of the sphere $x_1^2 + x_2^2 + x_3^2 + x_4^2 = R^2$ in $\mathbb{R}^4$ by (i) treating it as the union of two graphs $x_4 = \pm\sqrt{R^2 - x_1^2 - x_2^2 - x_3^2}$ and evaluating an integral in $x_1^2 + x_2^2 + x_3^2 \leq R^2$, and (ii) using the parametric representation in spherical polar coordinates

$$\begin{cases} x_1 &= R\sin\theta_2 \sin\theta_1 \cos\phi \\ x_2 &= R\sin\theta_2 \sin\theta_1 \sin\phi \\ x_3 &= R\sin\theta_2 \cos\theta_1 \\ x_4 &= R\cos\theta_2 \end{cases}$$

for $0 \leq \theta_1, \theta_2 \leq \pi, 0 \leq \phi \leq 2\pi$.

Hint. For (i), after a change of variables into spherical polar coordinates, it reduces to $4\pi R^3 \left(\int_0^1 \frac{\rho^2}{1-\rho^2}\, d\rho\right)$; for (ii), observe that the column vectors of the Jacobian matrix are orthogonal to each other.

4. Let $r = r(z) > 0$ be continuously differentiable on $[a, b]$. Then it generates a surface of revolution via the parametrization

$$(z, \theta) \mapsto (r(z)\cos\theta, r(z)\sin\theta, z) \in \mathbb{R}^3, a \leq z \leq b, 0 \leq \theta \leq 2\pi.$$

Verify that the area of this surface is given by

$$2\pi \int_a^b r(z)\sqrt{(r'(z))^2 + 1}\, dz.$$

$r_1(z) = \cosh(z)$ and $r_2(z) = \cosh(1)$ for $|z| \leq 1$ both can be used to generate a surface of revolution sharing the boundary $(\cosh(1)\cos(\theta), \cosh(1)\sin(\theta), \pm 1)$. which of them has less area?

6.7.4 The integral of a Function on an m-dimensional Differentiable Surface in $\mathbb{R}^n$

Suppose that $\mathbf{u} \in U \subset \mathbb{R}^m \mapsto F(\mathbf{u}) \in \mathbb{R}^n$ is a parametrization for a surface patch $F(U)$ and $f(\mathbf{x})$ is a function defined in a set in $\mathbb{R}^n$ which includes $F(U)$, then it is reasonable to define the **integral of** f **on** $F(U)$ by

$$\begin{aligned} \int_{F(U)} f(\mathbf{x})\, d\text{vol} &= \int_U f(F(\mathbf{u}))\sqrt{\det(DF(\mathbf{u})^{\mathrm{T}} DF(\mathbf{u}))}\, d\mathbf{u} \\ &= \int_U f(F(\mathbf{u}))\sqrt{\det(g_{ij}(\mathbf{u}))}\, d\mathbf{u} \end{aligned}$$

when this integral exists.

An analogous discussion as above shows that this definition is independent of the parametrization used. Thus we have reduced the integration of a function on a surface to the integration on a domain in the Euclidean space of a modified integrand. As a result, the usual properties of integrals such as the linearity in the integrand hold.

Example 6.7.11

For any continuous function $f(x,y,z)$ defined on the sphere $S_R : x^2+y^2+z^2 = R^2$ in $\mathbb{R}^3$, using cylindrical representation, its integral on S_R, $\int_{S_R} f(x,y,z)dA$ is given by

$$\int_0^{2\pi}\int_{-R}^{R} f(\sqrt{R^2-z^2}\cos\theta, \sqrt{R^2-z^2}\sin\theta, z)R\, dz\, d\theta.$$

For example,

$$\int_{S_R} x^2 dA = \int_0^{2\pi}\int_{-R}^{R}(R^2-z^2)\cos^2\theta R\, dz\, d\theta.$$

But this integral could also be evaluated using the symmetry of the sphere

$$\int_{S_R} x^2 dA = \int_{S_R} x^2 dA = \int_{S_R} x^2 dA = \int_{S_R}\frac{x^2+y^2+z^2}{3}dA = \frac{4R^4\pi}{3}.$$

Exercises

1. Evaluate $\int_{x^2+y^2=R^2, |z|\le h} x^2\, dA$.

2. Evaluate the integral $\int_S z\, dA$, where S is the region of the plane $x+y+z=1$ enclosed within the cylinder $x^2+y^2 \le 1$.

3. Evaluate the integral $\int_S z^2\, dA$, where S is the cylinder $x^2+y^2=1$ intersected between the planes $x+y+z=\pm 1$.

4. In the setting of Exercise 5.5.16, suppose that $h(\mathbf{x})$ is integrable in V, the integral $\int_V h(\mathbf{x})\, d\mathbf{x}$ can be evaluated using the parametrization $G : (\mathbf{x}', t) \in B(\mathbf{0}, r) \times (-\delta, \delta) \mapsto (\mathbf{x}', g(\mathbf{x}', t)) \in V$. Let $g_{ij}(\mathbf{x}', t) = D_i G(\mathbf{x}', t) \cdot D_j G(\mathbf{x}', t)$, where we interpret D_n as D_t.

(a) Verify that $\sqrt{\det(g_{ij}(\mathbf{x}', t))} = |\partial_t g(\mathbf{x}', t)|$.

(b) Verify that

$$|\partial_t g(\mathbf{x}', t)| \|Df(\mathbf{x}', g(\mathbf{x}', t))\| = \sqrt{1 + \sum_{i=1}^{n-1} |\partial_i g(\mathbf{x}', t)|^2}.$$

(c) Show that

$$\begin{aligned}\int_V h(\mathbf{x})\, d\mathbf{x} &= \int_{-\delta}^{\delta}\int_{B(\mathbf{0},r)} h(\mathbf{x}', g(\mathbf{x}', t))|\partial_t g(\mathbf{x}', t)|\, d\mathbf{x}' dt \\ &= \int_{-\delta}^{\delta}\int_{B(\mathbf{0},r)} \frac{h(\mathbf{x}', g(\mathbf{x}', t))}{\|Df(\mathbf{x}', g(\mathbf{x}', t))\|}\sqrt{1 + \sum_{i=1}^{n-1} |\partial_i g(\mathbf{x}', t)|^2}\, d\mathbf{x}' dt.\end{aligned}$$

Note that the integral

$$\int_{B(\mathbf{0},r)} \frac{h(\mathbf{x}', g(\mathbf{x}', t))}{\|Df(\mathbf{x}', g(\mathbf{x}', t))\|}\sqrt{1 + \sum_{i=1}^{n-1} |\partial_i g(\mathbf{x}', t)|^2}\, d\mathbf{x}' = \int_{V\cap\{f=t\}} \frac{h(\mathbf{x})}{\|Df(\mathbf{x})\|}\, dA$$

is an integral on the leaf $V \cap \{f = t\}$. This evaluation of the integral in terms of integrals on the leaves of level surfaces of some function is called **co-area formula**.

Here is a simple application of the co-area formula. To evaluate $\int_{x^2+y^2+z^2\leq R^2} h(x,y,z)\,dx\,dy\,dz$, we can choose $f(x,y,z) = \sqrt{x^2+y^2+z^2}$. Then the set $\{(x,y,z) : x^2 + y^2 + z^2 \leq R^2\}$ can be described as $\{(x,y,z) : 0 \leq f(x,y,z) \leq R\}$. Note that $\|Df(x,y,z)\| = 1$. Despite that $f(x,y,z)$ fails to be differentiable at $(x,y,z) = (0,0,0)$, we can still argue that

$$\int_{x^2+y^2+z^2\leq R^2} h(x,y,z)\,dx\,dy\,dz = \int_0^R \left(\int_{x^2+y^2+z^2=r^2} h(x,y,z)\,dA \right) dr.$$

Chapter 7

Exterior Differential Calculus and Integration of Differential Forms

Our next objective is to extend the notion of the curl and divergence of a vector field introduced in multi-variable calculus and the Green Theorem, the Stokes Theorem, and the Divergence Theorem to more general contexts.

7.1 A Brief Review of the Notion of Curl and Divergence of a Vector Field

In multi-variable calculus the line integral of a vector field $\vec{X} = (X_1(\mathbf{x}), \cdots, X_n(\mathbf{x})), n = 2$ or 3, along a path (or loop) $\Gamma : t \in [a, b] \mapsto \vec{r}(t)$, is defined as

$$\int_\Gamma \vec{X} \cdot d\vec{r} = \int_a^b (X_1(\vec{r}(t))r_1'(t) + \cdots + X_n(\vec{r}(t))r_n'(t))\, dt,$$

and the flux of $\vec{X}$ across a surface S, is defined as

$$\int_S \vec{X} \cdot \vec{n}\, dA,$$

where $\vec{n}(\mathbf{x})$ is a (continuous) choice of unit normal vector to S at $\mathbf{x}$.

One natural question is *what infinitesimal quantities measure the "strength" of the circulation along a closed loop or flux across a closed surface near a point?*

The answers turn out to be the **curl** and, respectively, **divergence** of the vector field $\vec{X}$ at the point. For $n = 3$, the curl of $\vec{X} = (X_1(\mathbf{x}), X_2(\mathbf{x}), X_3(\mathbf{x}))$ at $\mathbf{x}$ is the vector

$$\text{curl}\vec{X}(\mathbf{x}) = (\frac{\partial X_3}{\partial x_2} - \frac{\partial X_2}{\partial x_3}, \frac{\partial X_1}{\partial x_3} - \frac{\partial X_3}{\partial x_1}, \frac{\partial X_2}{\partial x_1} - \frac{\partial X_1}{\partial x_2}).$$

For $n = 2$, we can treat $\vec{X}$ as a special case of $n = 3$ with $X_3(\mathbf{x}) = 0$ and $X_i(\mathbf{x})$ for $i = 1, 2$ depend only on (x_1, x_2), so the curl of such a vector field takes the form of

$$(0, 0, \frac{\partial X_2}{\partial x_1} - \frac{\partial X_1}{\partial x_2}).$$

The divergence of $\vec{X} = (X_1(\mathbf{x}), X_2(\mathbf{x}), X_3(\mathbf{x}))$ is the scalar function

$$\text{div}\vec{X}(\mathbf{x}) = \sum_{j=1}^{3} \frac{\partial X_j(\mathbf{x})}{\partial x_j}.$$

The Stokes Theorem says that if $\Gamma : t \in [a, b] \mapsto \vec{r}(t)$ is a differentiable closed loop (meaning $\Gamma(a) = \Gamma(b)$) in $\mathbb{R}^3$ and spans a differentiable surface S, then

$$\int_\Gamma \vec{X} \cdot d\vec{r} = \int_S \text{curl}\vec{X}(\mathbf{x}) \cdot \vec{n}(\mathbf{x})\, dA,$$

where $\vec{n}(\mathbf{x})$ is an appropriately chosen unit normal vector field to S.

If we accept Stokes Theorem, it can be used to give some geometric interpretation for curl$\vec{X}$. Fix some $\mathbf{x}_0$ and a plane Π containing $\mathbf{x}_0$ with a unit normal vector $\vec{n}$. Take S to be the disc in the plane Π of radius $\epsilon > 0$ centered at $\mathbf{x}_0$ and Γ to be the boundary of this disc. Then

$$\lim_{\epsilon\to 0} \frac{1}{\pi\epsilon^2} \int_\Gamma \vec{X} \cdot d\vec{r} = \lim_{\epsilon\to 0} \frac{1}{\pi\epsilon^2} \int_S \text{curl}\vec{X}(\mathbf{x}) \cdot \vec{n}(\mathbf{x})\, dA = \text{curl}\vec{X}(\mathbf{x}_0) \cdot \vec{n}.$$

Thus curl$\vec{X}(\mathbf{x}_0) \cdot \vec{n}$ is the infinitesimal rate of circulation—circulation per unit area—of $\vec{X}$ along closed loops surrounding $\mathbf{x}_0$ in the plane through $\mathbf{x}_0$ with unit normal $\vec{n}$.

In fact, we can see why curl$\vec{X}$ is defined this way by examining

$$\lim_{\epsilon\to 0} \frac{1}{\pi\epsilon^2} \int_\Gamma \vec{X} \cdot d\vec{r}$$

without the knowledge of Stokes Theorem.

In the simple setting of the two dimensional plane, if $\vec{X}$ is continuously differentiable in a neighborhood of $\mathbf{x}_0 = (x_0, y_0)$, and we take Γ to be a small square loop of side length 2ϵ around $\mathbf{x}_0$, then

$$\begin{aligned}
&\int_\Gamma \vec{X} \cdot d\vec{r} \\
&= \int_{x_0-\epsilon}^{x_0+\epsilon} \{X_1(x, y_0 - \epsilon) - X_1(x, y_0 + \epsilon)\}\, dx + \int_{y_0-\epsilon}^{y_0+\epsilon} \{X_2(x_0 + \epsilon, y) - X_2(x_0 - \epsilon, y)\}\, dy \\
&= \int_{x_0-\epsilon}^{x_0+\epsilon} \int_{y_0-\epsilon}^{y_0+\epsilon} -\frac{\partial X_1(x, y)}{\partial y}\, dydx + \int_{y_0-\epsilon}^{y_0+\epsilon} \int_{x_0-\epsilon}^{x_0+\epsilon} \frac{\partial X_2(x, y)}{\partial x}\, dxdy \\
&= \int_{[x_0-\epsilon,x_0+\epsilon]\times[y_0-\epsilon,y_0+\epsilon]} \left\{\frac{\partial X_2(x, y)}{\partial x} - \frac{\partial X_1(x, y)}{\partial y}\right\} dxdy.
\end{aligned}$$

This is the simplest form of Green's theorem and motivates the definition of the curl of a two dimensional vector field $(X_1(x, y), X_2(x, y))$ as $\frac{\partial X_2(x,y)}{\partial x} - \frac{\partial X_1(x,y)}{\partial y}$. Note that at any point $\mathbf{x}_0 = (x_0, y_0)$, its value is determined as

$$\lim_{\epsilon\to 0} \frac{1}{\text{Area enclosed by } \Gamma} \int_\Gamma \vec{X} \cdot d\vec{r}.$$

In three dimension or higher, the simplest loops are planar ones, namely, a loop contained in the plane spanned by a pair of orthonormal vectors $\vec{\xi}$ and $\vec{\eta}$ --- we take $\mathbf{x}_0$ to be the origin for simplicity, and $\Gamma : t \mapsto \vec{\gamma}(t)$ be a closed loop near and enclosing $\mathbf{x}_0$ in the plane spanned by $\vec{\xi}$ and $\vec{\eta}$ given as

$$\vec{\gamma}(t) = x(t)\vec{\xi} + y(t)\vec{\eta}.$$

Assuming $\vec{X}$ to have the necessary differentiability, then Taylor expansion

$$\vec{X}(\vec{\gamma}(t)) = \vec{X}(\mathbf{x}_0) + \left[\frac{\partial \vec{X}}{\partial \mathbf{x}}(\mathbf{x}_0)\right](\vec{\gamma}(t) - \mathbf{x}_0) + h.o.t.\ (\|\vec{\gamma}(t) - \mathbf{x}_0\|)$$

gives the leading order term of $\int_\Gamma \vec{X} \cdot d\vec{r}$ to be

$$\begin{aligned}
&\sum_{i,j} \frac{\partial X_i(\mathbf{x}_0)}{\partial x_j} \int_\Gamma \left(x(t)\xi_j + y(t)\eta_j\right)\left(x'(t)\xi_i + y'(t)\eta_i\right) dt \\
=&\sum_{i,j} \frac{\partial X_i(\mathbf{x}_0)}{\partial x_j} \int_\Gamma \left(x(t)x'(t)\xi_i\xi_j + y(t)y'(t)\eta_i\eta_j + x(t)y'(t)\xi_j\eta_i + y(t)x'(t)\xi_i\eta_j\right) dt \\
=&\sum_{i,j} \frac{\partial X_i(\mathbf{x}_0)}{\partial x_j} \left(\xi_j\eta_i \int_\Gamma x(t)y'(t)dt + \xi_i\eta_j \int_\Gamma y(t)x'(t)dt\right)
\end{aligned}$$

where we have used that along any closed loop

$$\int_\Gamma \vec{X}(\mathbf{x}_0) \cdot \vec{\gamma}'(t)\, dt = \int_\Gamma \left[\frac{\partial \vec{X}}{\partial \mathbf{x}}(\mathbf{x}_0)\right] \mathbf{x}_0 \cdot \vec{\gamma}'(t)\, dt = 0,$$

and

$$\int_\Gamma x(t)x'(t)dt = \int_\Gamma y(t)y'(t)dt = 0,$$

but

$$\int_\Gamma x(t)y'(t)dt = -\int_\Gamma y(t)x'(t)dt$$

equals the area enclosed by Γ --- one may take Γ to be a square or circle loop to see this, so the leading order term of $\int_\Gamma \vec{X} \cdot d\vec{r}$ is

$$\sum_{i,j} \frac{\partial X_i(\mathbf{x}_0)}{\partial x_j} \left(\xi_j\eta_i - \xi_i\eta_j\right) \cdot (\text{Area enclosed by } \Gamma)\,.$$

In other words,

$$\lim_{\epsilon \to 0} \frac{1}{\text{Area enclosed by } \Gamma} \int_\Gamma \vec{X} \cdot d\vec{r} = \sum_{i,j} \frac{\partial X_i(\mathbf{x}_0)}{\partial x_j} \left(\xi_j\eta_i - \xi_i\eta_j\right).$$

This derivation works for any dimension n. We observe that this "infinitesimal strength" of circulation of $\vec{X}$ along loops near $\mathbf{x}_0$ depends on the above specific combinations of derivatives of $\vec{X}$ as well as on the plane in terms of a choice of an orthonormal basis; and the dependence on these derivatives of $\vec{X}$ and on $\vec{\xi}$ and $\vec{\eta}$ is linear in each when the remaining variables are held as constant —this is the notion of a **multi-linear** function; here, we momentarily relax the condition that $\vec{\xi}$ and $\vec{\eta}$ are orthonormal and allow them to be any pair of vectors. Furthermore, the dependence on $(\vec{\xi}, \vec{\eta})$ is antisymmetrical in $(\vec{\xi}, \vec{\eta})$, and if $\vec{\xi}'$ and $\vec{\eta}'$ are another pair of vectors such that $\text{Span}\{\vec{\xi}', \vec{\eta}'\} = \text{Span}\{\vec{\xi}, \vec{\eta}\}$, then the quantities $\xi_j'\eta_i' - \xi_i'\eta_j'$ and $\xi_j\eta_i - \xi_i\eta_j$ have a common proportionality constant equal to the determinant of the matrix that relates the two pairs of bases and equal ± 1 when both bases are orthonormal.

Exercise 7.1.1 Verify that if $[\vec{\xi}', \vec{\eta}'] = [\vec{\xi}, \vec{\eta}]A$ for some 2×2 matrix A, then $\xi_j'\eta_i' - \xi_i'\eta_j' = (\det A)(\xi_j\eta_i - \xi_i\eta_j)$, and that if $\{\vec{\xi}', \vec{\eta}'\}$ and $\{\vec{\xi}, \vec{\eta}\}$ are orthonormal, then A is an orthogonal matrix.

In the case of dimension 3, if we take

$$\vec{n} = \vec{\xi} \wedge \vec{\eta} = (\xi_2\eta_3 - \xi_3\eta_2, \xi_3\eta_1 - \xi_1\eta_3, \xi_1\eta_2 - \xi_2\eta_1),$$

then it is a unit normal to the plane spanned by $\vec{\xi}$ and $\vec{\eta}$, and

$$\begin{aligned}&\sum_{i,j} \frac{\partial X_i}{\partial x_j}(\mathbf{x}_0)\,(\xi_j\eta_i - \xi_i\eta_j)\\ =&(\frac{\partial X_3}{\partial x_2} - \frac{\partial X_2}{\partial x_3}, \frac{\partial X_1}{\partial x_3} - \frac{\partial X_3}{\partial x_1}, \frac{\partial X_2}{\partial x_1} - \frac{\partial X_1}{\partial x_2}) \cdot \vec{n},\end{aligned}$$

thus producing the concept of the *curl* of a vector field in dimension 3.

In the general dimension, using exterior algebra motivated by the above discussion and to be introduced soon, we see that $\xi_j\eta_i - \xi_i\eta_j$ are simply the components of $\vec{\xi} \wedge \vec{\eta}$, and

$$(\vec{\xi}, \vec{\eta}) \mapsto \sum_{i,j} \frac{\partial X_i}{\partial x_j}(\mathbf{x}_0)\,(\xi_j\eta_i - \xi_i\eta_j)$$

is a bilinear antisymmetric function on $(\vec{\xi}, \vec{\eta})$ (we may remove the orthonormal condition on $\vec{\xi}, \vec{\eta}$ now).

For any $i < j$, the coefficient of $(\xi_i\eta_j - \xi_j\eta_i)$ above is $\frac{\partial X_j}{\partial x_i} - \frac{\partial X_i}{\partial x_j}$. Thus the natural generalization of the notion of the curl of a vector field $\vec{X}$ in $\mathbb{R}^n, n > 3$, is not a vector field, but an object that acts on any pair of vectors in a bilinear and antisymmetric fashion. This is a heuristic reason for the notion of a **tensor**.

Another important question is *how do vector fields and their curls transform under a change of variables*? This is particularly important in the theory of *manifolds*, where there is usually no canonical coordinates to work with.

If $\mathbf{y} = T(\mathbf{x})$ is a continuously differentiable change of variables, then any continuously differentiable curve $\mathbf{x} = \vec{\gamma}(t)$ is mapped to a continuously differentiable curve $\mathbf{y} = T(\vec{\gamma}(t))$, with

$$\mathbf{y}'(t) = \left[\frac{\partial T_i(\mathbf{x})}{\partial x_j}\right] \Big|_{\mathbf{x}=\vec{\gamma}(t)} \mathbf{x}'(t).$$

Since the value of a vector field at any point is naturally identified to be the tangent vector of a continuously differentiable curve passing through that point, if $Y^i(T(\mathbf{x}))$ are the components of the vector field at $T(\mathbf{x})$ in the $\mathbf{y}$ coordinates that is transformed from $X(\mathbf{x})$ by T, we expect

$$Y(T(\mathbf{x})) = \begin{bmatrix} Y^1(T(\mathbf{x})) \\ \vdots \\ Y^n(T(\mathbf{x})) \end{bmatrix} = \left[\frac{\partial T_i(\mathbf{x})}{\partial x_j}\right] |_{\mathbf{x}=\vec{\gamma}(t)} \begin{bmatrix} X^1(\mathbf{x}) \\ \vdots \\ X^n(\mathbf{x}) \end{bmatrix}.$$

But the quantity

$$(X^1(\mathbf{x}), \ldots, X^n(\mathbf{x})) \cdot (x_1'(t), \ldots, x_n'(t)),$$

which is the integrand in $\int_\Gamma \vec{X} \cdot d\vec{r}$, would then transform to

$$Y(T(\mathbf{x}))^{\mathrm{T}} \left(\left[\frac{\partial T_i(\mathbf{x})}{\partial x_j}\right]^{-1}\right)^{\mathrm{T}} \left[\frac{\partial T_i(\mathbf{x})}{\partial x_j}\right]^{-1} \mathbf{y}'(t)$$

in the $\mathbf{y}$ coordinates. This is because $(X^1(\mathbf{x}), \ldots, X^n(\mathbf{x})) \cdot (x_1'(t), \ldots, x_n'(t))$ encodes the Euclidean inner product between vectors in the $\mathbf{x}$ coordinates, while the representation of this inner product in the $\mathbf{y}$ coordinates no longer has a simple form. This transformation is not only complicated, but makes it even harder to keep track of the relation between the components of the curl computed in the $\mathbf{y}$ and $\mathbf{x}$ coordinates.

It turns out that we get a much simpler resolution of this issue if we do not treat $X(\mathbf{x})$ as a vector field, but as a field of **covectors**, namely, at each point $\mathbf{x}$, $\mathbf{u} \mapsto \langle X(\mathbf{x}), \mathbf{u}\rangle := X(\mathbf{x}) \cdot \mathbf{u}$ is a linear function on the vector space of tangent vectors at that point. The simplification is due to the transformation laws of vectors and covectors, to be introduced next.

Exercise 7.1.2 Determine

$$\lim_{\epsilon\to 0} \frac{1}{\text{Area enclosed by } \Gamma} \int_\Gamma \vec{X} \cdot d\vec{r},$$

where Γ is the disc of radius ϵ centered at $\mathbf{0} \in \mathbb{R}^4$ of the plane

$$\begin{cases} x_1 \quad - ax_3 - cx_4 &= 0 \\ \quad x_2 - bx_3 - dx_4 &= 0 \end{cases}.$$

Hint. One could avoid finding explicitly an orthonormal basis for the plane by using Exercise 7.1.1.

7.2 Dual space, Tensor product, and Exterior Algebra

We are all familiar with representing a hyperplane in $\mathbb{R}^n$ in the form of

$$a_1x_1 + \ldots + a_nx_n = b \text{ for some } a_1, \ldots, a_n, \text{ not all zero, and some } b.$$

Restricting to the case $b = 0$, such a hyperplane can be thought of as the *null space* of the **linear function**

$$\alpha(\mathbf{x}) := a_1x_1 + \ldots + a_nx_n$$

defined on the vectors $\mathbf{x} \in \mathbb{R}^n$.

Definition 7.2.1 Covectors and Dual Space.

The set of all linear functions on a (finite dimensional) vector space V is called the dual space of V, and is denoted as V^*. Elements of V^* are called covectors.

A multilinear function of order m if a function $\alpha : V \times \cdots \times V \mapsto \mathbb{R}$ (with m factors of V) such that $\alpha(\mathbf{x}_1, \cdots, \mathbf{x}_m)$ is linear in each $\mathbf{x}_i$ when the rest is held as fixed. Such an α is also called a covariant tensor of order m. The space of covariant tensors of order m on V is denoted as $\mathcal{T}^m(V)$.

Each non-zero element in V^* determines a hyperplane (through the origin) in V, namely, a subspace of codimension one; and two such elements determine the same hyperplane, if they are non-zero multiple of each other.

There are at least two ways to describe a subspace of V of codimension bigger than one. One way is as the intersection of several codimension one hypersurfaces: $\{\mathbf{x} \in V : \alpha_i(\mathbf{x}) = 0, 1 \le i \le m\}$, namely, as the set of solutions of a system of m linear homogeneous equations. If $\dim \operatorname{Span}\{\alpha_i, 1 \le i \le m\} = l$, then this is a codimension l subspace. This is seen by writing out each α_i in terms of its coefficients, then $\{\mathbf{x} \in V : \alpha_i(\mathbf{x}) = 0, 1 \leq i \leq m\}$ is the solution set of a system of m linear equations in n variables, with a coefficient matrix who row rank is l, thus the solution space is $n - l$ dimensional.

Another way is to choose a basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ for a subspace of V. But there are other choices for a basis of this subspace. Exterior algebra provides a convenient tool to identify a subspace (with orientation). Suppose $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is another basis of the subspace, then we can write

$$\mathbf{v}_j = \sum_{i=1}^{k} a_{ij}\mathbf{u}_i, 1 \leq j \leq k \tag{7.2.1}$$

for some coefficients a_{ij}. This would form a $k \times k$ invertible matrix A with a_{ij} as its entries. (7.2.1) can be written compactly in a matrix form:

$$[\mathbf{v}_1 \cdots \mathbf{v}_k] = [\mathbf{u}_1 \cdots \mathbf{u}_k]A. \tag{7.2.2}$$

It turns out that the following product between vectors, a generalization of the cross product between vectors in $\mathbb{R}^3$ called the **exterior product**, provides an efficient tool to describe oriented subspaces of V. We first give a preliminary formal definition to illustrate its usage; a more precise definition, together with the justification for the existence/construction of this product, will be given shortly.

Definition 7.2.2 Exterior Product.

Let V be a vector space. The exterior product is a product $\mathbf{u} \wedge \mathbf{v}$ between $\mathbf{u}, \mathbf{v} \in V$ such that it is linear in each factor and antisymmetric in $\mathbf{u}, \mathbf{v}$:

$$\mathbf{u} \wedge \mathbf{v} = -\mathbf{v} \wedge \mathbf{u}.$$

Furthermore, this product extends to any k-tuple of vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ in V which obeys associativity and is linear in each factor and antisymmetric in any adjacent pairs.

As an illustration, take $k = 3$, then

$$(\mathbf{u}_1 \wedge \mathbf{u}_2) \wedge \mathbf{u}_3 = \mathbf{u}_1 \wedge (\mathbf{u}_2 \wedge \mathbf{u}_3) = -\mathbf{u}_2 \wedge \mathbf{u}_1 \wedge \mathbf{u}_3 = -\mathbf{u}_1 \wedge \mathbf{u}_3 \wedge \mathbf{u}_2.$$

Back to (7.2.2). The algebraic rules of exterior algebra would lead to

$$\mathbf{v}_1 \wedge \ldots \wedge \mathbf{v}_k = (\det A)\, \mathbf{u}_1 \wedge \ldots \wedge \mathbf{u}_k. \tag{7.2.3}$$

This is particularly easy to see in the case of $k = 2$:

$$\mathbf{v}_1 \wedge \mathbf{v}_2 = (a_{11}\mathbf{u}_1 + a_{21}\mathbf{u}_2) \wedge (a_{12}\mathbf{u}_1 + a_{22}\mathbf{u}_2) = (a_{11}a_{22} - a_{12}a_{21})\, \mathbf{u}_1 \wedge \mathbf{u}_2.$$

Thus the exterior product of a basis $\mathbf{v}_1 \wedge \ldots \wedge \mathbf{v}_k$, up to the scaling factor $\det A$, is independent of the choice of a basis for the subspace. In fact, the sign of $\det A$ can be used to identify whether the two bases $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ and $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ are in the same or opposite **orientation** of the subspace.

Proposition 7.2.3

For any basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ of a vector space V, there is a unique basis $\{\alpha_1, \ldots, \alpha_n\}$ of V^* such that

$$\alpha_i(\mathbf{u}_j) = \delta_{ij} \text{ for } 1 \leq i, j \leq n. \tag{7.2.4}$$

Conversely, for any basis $\{\alpha_1, \ldots, \alpha_n\}$ *of* V^**, there is a unique basis* $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ *of* V *such that (7.2.4) holds.*

Definition 7.2.4 Dual Basis.

$\{\alpha_1, \ldots, \alpha_n\}$ above is called the dual basis of $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$, and the latter is called the dual basis of $\{\alpha_1, \ldots, \alpha_n\}$.

Definition 7.2.5 Inner Product on a Vector Space.

An inner product on a vector space V is a symmetric positive definite covariant tensor of order 2, namely, a bilinear and symmetric function g on V such that $g(\mathbf{x}, \mathbf{x}) > 0$ unless $\mathbf{x} = \mathbf{0}$. Such a g is also called a metric.

On an inner product space one can define orthonormal bases. If two bases $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ and $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ are orthonormal with respect to a metric g, and are related via (7.2.1), then the matrix A must be an orthogonal matrix.

Definition 7.2.6 Isometry of a Metric.

A linear map $T : V \mapsto V$ is called an isometry of g, if $g(T\mathbf{x}, T\mathbf{x}) = g(\mathbf{x}, \mathbf{x})$ holds for all $\mathbf{x} \in V$.

This condition is equivalent to $g(T\mathbf{x}, T\mathbf{y}) = g(\mathbf{x}, \mathbf{y})$ holds for all $\mathbf{x}, \mathbf{y} \in V$.

On an inner product space V with g as its inner product, there is an isomorphism $\sharp : V^* \mapsto V$ such that

$$\alpha(\mathbf{x}) = g(\sharp\alpha, \mathbf{x}) \text{ for } \alpha \in V^*, \mathbf{x} \in V.$$

This induces an inner product on V^* via

$$g(\alpha, \beta) = g(\sharp\alpha, \sharp\beta) \text{ for } \alpha, \beta \in V^*.$$

An abstract vector space does not have a natural definition of volume even for cells of the form $\{s_1\mathbf{u}_1 + \ldots + s_k\mathbf{u}_k : 0 \le s_i \le 1, 1 \le i \le k\}$. But once an inner product g is introduced, it is natural to define the k dimensional volume of such cells to be 1 whenever $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is an orthonormal basis. This volume is invariant under all isometries of g. If $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is an orthonormal basis of g and (7.2.1) holds, then (7.2.3) shows that the corresponding cell generated by $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ has volume equal to $|\det A|$; in other words, the exterior vector $\mathbf{v}_1 \wedge \ldots \wedge \mathbf{v}_k$ encodes both the volume and orientation of the parallelepiped formed with these vectors as edges.

An inner product on V is just a special kind of bilinear function on V.

Definition 7.2.7 Tensor Product.

The tensor product of any two linear functions α, β on V is the bilinear function

$$\alpha \otimes \beta(\mathbf{x}, \mathbf{y}) = \alpha(\mathbf{x})\beta(\mathbf{y}) \text{ for } \mathbf{x}, \mathbf{y} \in V.$$

If $\alpha \in \mathcal{T}^m(V)$ and $\beta \in \mathcal{T}^l(V)$, then $\alpha \otimes \beta \in \mathcal{T}^{m+l}(V)$ is defined by

$$\begin{aligned} &\alpha \otimes \beta(\mathbf{x}_1, \ldots, \mathbf{x}_m, \mathbf{x}_{m+1}, \ldots, \mathbf{x}_{m+l}) \\ =&\alpha(\mathbf{x}_1, \ldots, \mathbf{x}_m)\beta(\mathbf{x}_{m+1}, \ldots, \mathbf{x}_{m+l}) \text{ for } \mathbf{x}_1, \ldots, \mathbf{x}_m, \mathbf{x}_{m+1}, \ldots, \mathbf{x}_{m+l} \in V. \end{aligned}$$

Note that $\alpha \otimes \beta \neq \beta \otimes \alpha$ in general.

For any basis $\{\alpha_1, \dots, \alpha_n\}$ of V^* for $m \geq 2$, $\sum_{i=1}^n \alpha_i \otimes \alpha_i$ defines an inner product on V so that its dual basis is an orthonormal basis in this metric.

Definition 7.2.8 Alternating Form.

An m-linear function $\omega \in \mathcal{T}^m(V)$ is called an alternating form on V if

$$\omega(\mathbf{x}_{\sigma(1)}, \dots, \mathbf{x}_{\sigma(m)}) = \operatorname{sgn} \sigma \, \omega(\mathbf{x}_1, \dots, \mathbf{x}_m) \text{ for all permutations } \sigma.$$

The subspace of alternating forms on V is denoted as $\Lambda^m(V)$.

Definition 7.2.9 Alt Map.

We define $\text{Alt} : \mathcal{T}^m(V) \mapsto \Lambda^m(V)$ by

$$\text{Alt}(\omega)(\mathbf{x}_1, \dots, \mathbf{x}_m) = \frac{1}{m!} \sum_\sigma \operatorname{sgn} \sigma \, \omega(\mathbf{x}_{\sigma(1)}, \dots, \mathbf{x}_{\sigma(m)}),$$

and define the **wedge product**, also called **exterior product**,

$$\alpha \wedge \beta = \frac{(m+l)!}{m!\, l!} \text{Alt}(\alpha \otimes \beta) \text{ for } \alpha \in \Lambda^m(V), \beta \in \Lambda^l(V).$$

Note that

$$\text{Alt}(\omega) = \omega \text{ if } \omega \in \Lambda^m(V).$$

The following property will be used often.

$$(\alpha \wedge \beta) \wedge \gamma = \alpha \wedge (\beta \wedge \gamma) = \frac{(k+l+m)!}{k!\, l!\, m!} \text{Alt}(\alpha \otimes \beta \otimes \gamma) \tag{7.2.5}$$

for $\alpha \in \Lambda^m(V), \beta \in \Lambda^l(V), \gamma \in \Lambda^k(V)$.

In the case of $m = l = 1$,

$$\text{Alt}(\alpha \otimes \beta) = \frac{1}{2} (\alpha \otimes \beta - \beta \otimes \alpha),$$

and

$$\alpha \wedge \beta = \alpha \otimes \beta - \beta \otimes \alpha.$$

In the case of $k = l = m = 1$ we also get

$$\begin{aligned} &(\alpha \wedge \beta) \wedge \gamma \\ =&\alpha \otimes \beta \otimes \gamma + \beta \otimes \gamma \otimes \alpha + \gamma \otimes \alpha \otimes \beta \\ &- \beta \otimes \alpha \otimes \gamma - \alpha \otimes \gamma \otimes \beta - \gamma \otimes \beta \otimes \alpha. \end{aligned}$$

It also follows that $\alpha \wedge \alpha = 0$ if $\alpha \in \Lambda^1(V)$. However, if $n \geq 4$ and $\{\alpha_1, \dots, \alpha_n\}$ is a basis of V^*, then

$$(\alpha_1 \wedge \alpha_2 + \alpha_3 \wedge \alpha_4) \wedge (\alpha_1 \wedge \alpha_2 + \alpha_3 \wedge \alpha_4) = 2\alpha_1 \wedge \alpha_2 \wedge \alpha_3 \wedge \alpha_4 \neq 0.$$

If $\{\alpha_1, \dots, \alpha_n\}$ is a basis of V^*, then $\{\alpha_i \otimes \alpha_j : 1 \leq i, j \leq n\}$ forms a basis of $\mathcal{T}^2(V)$, and $\{\alpha_i \wedge \alpha_j : 1 \leq i < j \leq n\}$ forms a basis of $\Lambda^2(V)$. $\{\alpha_i \otimes \alpha_j + \alpha_j \otimes \alpha_i : 1 \leq i \leq j \leq n\}$ forms a basis of $\mathcal{S}^2(V)$, the space of symmetric two tensors of V.

Alternatively, a tensor $g \in \mathcal{S}^2(V)$ is determined by its actions on $\{(\mathbf{u}_i, \mathbf{u}_j) : 1 \leq i \leq j \leq n\}$, while a tensor $\omega \in \Lambda^2(V)$ is determined by its actions on $\{(\mathbf{u}_i, \mathbf{u}_j) : 1 \leq i < j \leq n\}$. Here $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ is the dual basis of $\{\alpha_1, \dots, \alpha_n\}$.

For any $\mathbf{x} = \sum_{i=1}^n x_i \mathbf{u}_i$ and $\mathbf{y} = \sum_{i=1}^n y_i \mathbf{u}_i$, then a symmetric 2-tensor g satisfies

$$\begin{aligned} g(\mathbf{x}, \mathbf{y}) =& g(\sum_{i=1}^n x_i \mathbf{u}_i, \sum_{i=1}^n y_i \mathbf{u}_i) \\ =& \sum_{i=1}^n \sum_{j=1}^n x_i y_j g(\mathbf{u}_i, \mathbf{u}_j) \\ =& \sum_{i=1}^n x_i y_i g(\mathbf{u}_i, \mathbf{u}_i) + \sum_{i<j} (x_i y_j + x_j y_i)\, g(\mathbf{u}_i, \mathbf{u}_j) \end{aligned}$$

while an antisymmetric 2-tensor ω satisfies

$$\omega(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n \sum_{j=1}^n x_i y_j \omega(\mathbf{u}_i, \mathbf{u}_j) = \sum_{i<j} (x_i y_j - x_j y_i)\, \omega(\mathbf{u}_i, \mathbf{u}_j).$$

For example, when $n = 2$, $\mathcal{S}^2(\mathbb{R}^2)$ is three dimensional, while $\Lambda^2(\mathbb{R}^2)$ is one dimensional, and $g \in \mathcal{S}^2(\mathbb{R}^2)$ is determined by

$$g(\mathbf{x}, \mathbf{y}) = x_1 y_1 g(\mathbf{u}_1, \mathbf{u}_1) + (x_1 y_2 + x_2 y_1) g(\mathbf{u}_1, \mathbf{u}_2) + x_2 y_2 g(\mathbf{u}_2, \mathbf{u}_2),$$

while $\omega \in \Lambda^2(\mathbb{R}^2)$ is determined by

$$\omega(\mathbf{x}, \mathbf{y}) = (x_1 y_2 - x_2 y_1) \omega(\mathbf{u}_1, \mathbf{u}_2).$$

In computations sometimes $\alpha_i \wedge \alpha_j$ may show up even when $i \geq j$, but to identify the coefficients of the resulting alternating tensor, one needs to transform all terms in terms of the basis discussed above. For example, $a\, \alpha \otimes \beta + b\, \beta \otimes \alpha \in \mathcal{T}^2(V)$, and if we apply the Alt operation on it, we get a tensor in $\Lambda^2(V)$

$$\frac{a}{2} (\alpha \otimes \beta - \beta \otimes \alpha) + \frac{b}{2} (\beta \otimes \alpha - \alpha \otimes \beta),$$

which could be recognized to be $\frac{a}{2} \alpha \wedge \beta + \frac{b}{2} \beta \wedge \alpha$ but ends up identified as $\frac{a-b}{2} \alpha \wedge \beta$. The discussion here applies to higher order tensors as well.

We can treat vectors in V as linear functions on V^*, then tensor product and exterior product on V make sense. For instance, for any $\mathbf{u}, \mathbf{v} \in V$, $\mathbf{u} \otimes \mathbf{v} \in \mathcal{T}^2(V^*)$ in the sense that $\mathbf{u} \otimes \mathbf{v}(\alpha, \beta) = \alpha(\mathbf{u})\beta(\mathbf{v})$ and $\mathbf{u} \wedge \mathbf{v} \in \Lambda^2(V^*)$ in the sense that $\mathbf{u} \wedge \mathbf{v}(\alpha, \beta) = \alpha(\mathbf{u})\beta(\mathbf{v}) - \alpha(\mathbf{v})\beta(\mathbf{u})$.

For any linear transformation $L : V \mapsto W$, there is a naturally defined adjoint map, labeled as L^*, such that for any $\alpha \in W^*$,

$$L^*(\alpha) \in V^* \text{ is defined via } L^*(\alpha)(\mathbf{x}) = \alpha(L(\mathbf{x})) \text{ for all } \mathbf{x} \in V. \tag{7.2.6}$$

In fact, one can define L^* on $\mathcal{T}^m(W)$ in a similar way such that for any $\omega \in \mathcal{T}^m(W)$,

$$L^*(\omega)(\mathbf{x}_1, \ldots, \mathbf{x}_m) = \omega(L(\mathbf{x}_1), \ldots, L(\mathbf{x}_m)) \text{ for all } \mathbf{x}_1, \ldots, \mathbf{x}_m \in V.$$

For a metric g on W, $L^*(g)$ is a metric on V provided that L is injective. In such a case, we call $L^*(g)$ the **pull-back metric** of g by L.

When ω is an alternating tensor on W, $L^*(\omega)$ is an alternating tensor on V. Furthermore, for two alternating tensors α, β on W,

$$L^*(\alpha \wedge \beta) = L^*(\alpha) \wedge L^*(\beta).$$

A property related to (7.2.6) is that for any two bases $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ and $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ of V, let $\{\alpha_1, \cdots, \alpha_n\}$ and $\{\beta_1, \cdots, \beta_n\}$ be their respective dual bases in V^*, then for any vector $X \in V$, and covector $\omega \in V^*$, if

$$X = \sum_{i=1}^n x_i \mathbf{u}_i = \sum_{i=1}^n y_i \mathbf{v}_i, \quad \omega = \sum_{i=1}^n a_i \alpha_i = \sum_{i=1}^n b_i \beta_i$$

then

$$\omega(X) = \sum_{i=1}^{n} a_i x_i = \sum_{i=1}^{n} b_i y_i.$$

In the context of Stokes Theorem we will treat $\sum_{i=1}^{n} X_i(\mathbf{x})x_i'(t)$ as the pairing between a vector and a covector and will apply the above transformation property when applying a change of variables.

Exercises

1. Let $\alpha, \beta \in V^*$ be such that $\{\mathbf{x} \in V : \alpha(\mathbf{x}) = 0\}$ is identical to $\{\mathbf{x} \in V : \beta(\mathbf{x}) = 0\}$. Prove that there exists some constant c such that $\alpha = c\beta$.

2. Prove that $T : V \mapsto V$ is an isometry with respect to the metric g iff $g(T\mathbf{x}, T\mathbf{y}) = g(\mathbf{x}, \mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in V$.

3. Let $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ be a basis of V and $\{\alpha_1, \cdots, \alpha_n\}$ be its dual basis. Let g be a metric on V, and $g_{ij} = g(\mathbf{u}_i, \mathbf{u}_j)$. Then for the covector $\alpha = \sum_{i=1}^{n} a_i \alpha_i \in V^*$, we have $\sharp\alpha = \sum_{i,j=1}^{n} a_i g^{ij} \mathbf{u}_j$, where g^{ij} are the coefficients of the inverse matrix of $[g_{ij}]$.

4. Prove the general case of (7.2.3).

5. Let $\{\alpha_1, \ldots, \alpha_{2n}\}$ be a basis of V^* and $\omega = \alpha_1 \wedge \alpha_2 + \alpha_3 \wedge \alpha_4 + \cdots + \alpha_{2n-1} \wedge \alpha_{2n} \in \Lambda^2(V)$. Prove that

$$\omega \wedge \cdots \wedge \omega = n!\ \alpha_1 \wedge \alpha_2 \wedge \alpha_3 \wedge \alpha_4 \wedge \cdots \wedge \alpha_{2n-1} \wedge \alpha_{2n},$$

where the wedge product has n factors of ω.

6. Let $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ be a basis of V and $\{\alpha_1, \cdots, \alpha_n\}$ be its dual basis in V^*, $\{\mathbf{v}_1, \cdots, \mathbf{v}_m\}$ a basis of W and $\{\beta_1, \cdots, \beta_m\}$ be its dual basis in W^*. Let $L : V \mapsto W$ be a linear map and the $m \times n$ matrix A be the matrix representation of L with respect to the bases $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ and $\{\mathbf{v}_1, \cdots, \mathbf{v}_m\}$. Then L^* is represented by A^{T} with respect to the bases $\{\beta_1, \cdots, \beta_m\}$ and $\{\alpha_1, \cdots, \alpha_n\}$.

7. Let $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ and $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ of V be two bases of V, and $\{\alpha_1, \cdots, \alpha_n\}$ and $\{\beta_1, \cdots, \beta_n\}$ be their respective dual bases in V^*. Suppose that $\mathbf{v}_i = \sum_{k=1}^{n} a_{ik} \mathbf{u}_k$ for some matrix $A = [a_{ik}]$. Prove that $\beta_i = \sum_{k=1}^{n} b_{ik} \alpha_k$, where $[b_{ik}] = (A^{-1})^T$.

8. Prove (7.2.5) and use it to show that for any k covectors $\{\alpha_1, \cdots, \alpha_k\}$ in V^* and k vectors $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ in V,

$$\alpha_1 \wedge \cdots \wedge \alpha_k(\mathbf{u}_1, \cdots, \mathbf{u}_k) = \det\left[\alpha_i(\mathbf{u}_j)\right].$$

7.3 Vector Fields and Differential Forms

7.3.1 Definition and Basic Properties of Vector Fields and Differential One-Forms

We start with an initial definition of a vector field in a Euclidean domain; it is to be modified later to be suitable in a more general context such as on a differentiable surface.

Definition 7.3.1 Vector Field in a Euclidean Domain.

Let $U \subset \mathbb{R}^n$ be open. A vector field in U is an $\mathbb{R}^n$-valued function $X : \mathbf{x} \in U \mapsto \mathbb{R}^n$. A vector field $X(\mathbf{x})$ is continuous (differentiable) in U if $X(\mathbf{x})$ as an $\mathbb{R}^n$-valued function is continuous (differentiable) in U.

We will use a vector field mostly in the operation of taking directional derivatives of differentiable functions. First, recall the directional derivative of a differentiable function f in $\mathbb{R}^n$ in the direction $\mathbf{v}$ at a point P:

$$D_{\mathbf{v}} f(P) = \frac{df(P + t\mathbf{v})}{dt}\Big|_{t=0} = \sum_{i=1}^{n} v^i \frac{\partial f}{\partial x_i}(P),$$

where $\mathbf{v} = (v^1, \ldots, v^n)$. This relation holds for any differentiable curve $\vec{\gamma}(t)$ passing through P at $t = 0$ and $\vec{\gamma}'(0) = \mathbf{v}$:

$$\frac{d(f \circ \gamma)}{dt}\Big|_{t=0} = D_{\mathbf{v}} f(P).$$

$\mathbf{v}$ is called a **tangent vector** at P. The set of tangent vectors at P forms a vector space, called the **tangent space** of $\mathbb{R}^n$ at P, and is denoted as $\mathbb{R}^n_P$. For the situation here, $\mathbb{R}^n_P$ is simply a copy of $\mathbb{R}^n$, and $\mathbb{R}^n_P$ at different P are identified with each other in a natural way. But we will see that this is not the case when we discuss the tangent space of a differentiable surface, as there is no obvious natural relations between vectors at different points, and it is important to associate a tangent vector to a specific point.

Suppose that $X(\mathbf{x})$ is a continuous vector field on U and $f(\mathbf{x})$ is a continuously differentiable function on U, then at each $\mathbf{x} \in U$, $X(\mathbf{x})$ is a tangent vector at $\mathbb{R}^n_{\mathbf{x}}$ and $D_{X(\mathbf{x})} f(\mathbf{x})$ is a continuous function on U.

The discussions in the previous paragraphs generalize to a (differentiable) surface in $\mathbb{R}^n$. We will first work with a patch of a differentiable surface as given by a differentiable map $\vec{\gamma}(\mathbf{u}) : U \mapsto \mathbb{R}^n$ defined for the parameter $\mathbf{u} \in U \subset \mathbb{R}^k$ for some open domain U. A differentiable curve on the surface through $P_0 = \vec{\gamma}(\mathbf{u}_0)$ is given in terms of $\vec{\gamma}(\mathbf{u}(t))$, where $\mathbf{u}(t)$ is a differentiable curve in the parameter domain U with $\mathbf{u}(0) = \mathbf{u}_0$. Then the chain rule

$$[\vec{\gamma}(\mathbf{u}(t))]' = D_{\mathbf{u}}\vec{\gamma}(\mathbf{u}(t))\mathbf{u}'(t)$$

implies that the tangent of the curve $\vec{\gamma}(\mathbf{u}(t))$ at P_0, $[\vec{\gamma}(\mathbf{u}(t))]'|_{t=0}$, is a linear combination of $D_{u_i}\vec{\gamma}(\mathbf{u}_0)$, where $D_{u_i}\vec{\gamma}(\mathbf{u}_0)$ is an alternative notation for $D_{\mathbf{e}_i}\vec{\gamma}(\mathbf{u}_0)$, namely, the partial derivative of $\vec{\gamma}(\mathbf{u})$ with respect to its ith coordinate. The span of these k vectors forms the tangent space of the surface at P_0.

In order for such a parametrization $\vec{\gamma}(\mathbf{u})$ to represent a geometric k-dimensional surface, we require that these k vectors be linearly independent so the tangent space to the surface is k-dimensional, namely the matrix $[D_{u_1}\vec{\gamma}(\mathbf{u}_0) \ldots D_{u_k}\vec{\gamma}(\mathbf{u}_0)]$ has rank k. Let S denote this patch of differentiable surface. Then there is a well defined tangent space $S_{\vec{\gamma}(\mathbf{u})}$ for every $\mathbf{u}$. When $\vec{\gamma}(\mathbf{u})$ is continuously differentiable, we have a sense that the tangent space $S_{\vec{\gamma}(\mathbf{u})}$ varies continuously with $\mathbf{u}$. But that is not a topic to be taken up now. Instead, we first focus on how a tangent vector is used to compute the directional derivative of a differentiable function.

If f is a differentiable function defined on $\mathbb{R}^n$, then its restriction to the surface S becomes a function on the surface. To consider its differentiability properties on the surface, we use the parametrization $\vec{\gamma}(\mathbf{u})$ to get a function $f \circ \vec{\gamma}$ in the parameter domain of $\mathbf{u}$, then its directional derivative at P_0 in the direction of $D_{u_i}\vec{\gamma}(\mathbf{u}_0)$ is given by

$$D_{D_{u_i}\vec{\gamma}(\mathbf{u}_0)}f\Big|_{\vec{\gamma}(\mathbf{u}_0)} = \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(\vec{\gamma}(\mathbf{u}_0))D_{\mathbf{e}_i}\vec{\gamma}_j(\mathbf{u}_0) = \frac{\partial\,(f\circ\vec{\gamma})}{\partial u_i}(\mathbf{u}_0).$$

It is this kind of consideration that makes it natural to use $\frac{\partial}{\partial u_i}\Big|_{\vec{\gamma}(\mathbf{u}_0)}$ to represent the tangent vector $D_{u_i}\vec{\gamma}(\mathbf{u}_0)$ to the surface at $P_0 = \vec{\gamma}(\mathbf{u}_0)$, when $\vec{\gamma}(\mathbf{u})$ is a parametric representation of the surface. Namely,

- $\frac{\partial\vec{\gamma}}{\partial u_i} = D_{u_i}\vec{\gamma}$ is a geometric tangent vector to the surface $\mathbf{x} = \vec{\gamma}(\mathbf{u})$ at $\vec{\gamma}(\mathbf{u})$ arising from a curve whose parametrization in the $\mathbf{u}$ parameter space runs along the u_i direction.
- $\frac{\partial(f\circ\vec{\gamma})}{\partial u_i} = D_{D_{u_i}\vec{\gamma}}f$ means that, as an operator, $\frac{\partial}{\partial u_i}$ represents directional derivative in the direction of the tangent $D_{u_i}\vec{\gamma}$.

Thus

$$D_{D_{u_i}\vec{\gamma}(\mathbf{u}_0)}f\Big|_{\vec{\gamma}(\mathbf{u}_0)} = D_{\frac{\partial}{\partial u_i}\Big|_{\vec{\gamma}(\mathbf{u}_0)}}f.$$

In this notation, $\left\{\frac{\partial}{\partial u_1}\Big|_P, \ldots, \frac{\partial}{\partial u_k}\Big|_P\right\}$ forms a basis of S_P. The advantage of this notation will become clear when a change of variable is used, which would cause a change of basis for the tangent space. We can write any $\mathbf{v} \in S_P$ as $\mathbf{v} = \sum_{i=1}^{k} v^i \frac{\partial}{\partial u_i}\Big|_P$, then

$$D_{\mathbf{v}}f(P) = \sum_{i=1}^{k} v^i \frac{\partial(f\circ\vec{\gamma})}{\partial u_i}(\mathbf{u})$$

for $P = \vec{\gamma}(\mathbf{u})$.

Definition 7.3.2 Vector Field on a Differentiable Surface.

A vector field X on a differentiable surface S is a map $P \in S \mapsto X(P) \in S_P$. Namely, it assigns to each $P \in S$ a tangent vector $X(P)$ to S at P.

When S is given by a differentiable parametrization $\vec{\gamma} : U \subset \mathbb{R}^k \mapsto S$, any vector field X on S can be represented as $X(\vec{\gamma}(\mathbf{u})) = \sum_{i=1}^{k} X_i(\mathbf{u})\frac{\partial}{\partial u_i}\Big|_{\vec{\gamma}(\mathbf{u})}$. X is said to be a continuous (or continuously differentiable) vector field on S if the coefficient functions $X_i(\mathbf{u})$ are continuous (or continuously differentiable) functions of $\mathbf{u} \in U$.

We will often write $\frac{\partial}{\partial u_i}$ for $\frac{\partial}{\partial u_i}\Big|_{\vec{\gamma}(\mathbf{u})}$ to simplify notations. Note that $\frac{\partial}{\partial u_i}$ is a vector field on S, but if the u_i's are not used in connection with the parametrization, then $\frac{\partial}{\partial u_i}$ also represents a vector field in U, which takes the vector $\mathbf{e}_i$ everywhere in U. One should watch out for the context in which this notation is used. In the latter context, a vector field in U is simply an R^k-valued function, so one can take its derivative and in this case $D_{\mathbf{v}}\left(\frac{\partial}{\partial u_i}\right) = 0$ for any $\mathbf{v} \in \mathbb{R}^k_{\mathbf{u}}$. But in the former context, $\mathbf{v}$ should be regarded as the tangent vector $\frac{d\vec{\gamma}(\mathbf{u}+t\mathbf{v})}{dt}\Big|_{t=0} = D\vec{\gamma}(\mathbf{u})\mathbf{v} \in S_{\vec{\gamma}(\mathbf{u})}$ on S, and $D_{\mathbf{v}}\left(\frac{\partial}{\partial u_i}\right)$ should be related to $D_{D\vec{\gamma}(\mathbf{u})\mathbf{v}}\left(\frac{\partial\vec{\gamma}(\mathbf{u})}{\partial u_i}\right)$. But this output in $\mathbb{R}^n$ may not be a tangent vector to S at $\vec{\gamma}(\mathbf{u})$. There is a way to obtain a tangent vector to S at $\vec{\gamma}(\mathbf{u})$ through orthogonal projection. This will introduce the notion of **covariant differentiation** of vector field on S. But we will not pursue that topic in this course.

Example 7.3.3 Vector fields on a graph.

A graph G of a differentiable function $h(\mathbf{u})$ for $\mathbf{u} \in \mathbb{R}^{n-1}$ can be parametrized as $\vec{\gamma}(\mathbf{u}) = (\mathbf{u}, h(\mathbf{u}))$. Then the vector field $\frac{\partial}{\partial u_i}$ is simply a coordinate representation for the geometric vector field $(\mathbf{e}_i, \frac{\partial h(\mathbf{u})}{\partial u_i})$ on G, and

$$\begin{aligned} D_{\frac{\partial}{\partial u_i}} f(\mathbf{u}, h(\mathbf{u})) = D_{(\mathbf{e}_i, \frac{\partial h(\mathbf{u})}{\partial u_i})} f(\mathbf{u}, h(\mathbf{u})) =& \frac{\partial f(\mathbf{u}, h(\mathbf{u}))}{\partial u_i} \\ =& \frac{\partial f}{\partial x_i}(\mathbf{u}, h(\mathbf{u})) + \frac{\partial f}{\partial x_n}(\mathbf{u}, h(\mathbf{u}))\frac{\partial h(\mathbf{u})}{\partial u_i}. \end{aligned}$$

$\{\frac{\partial}{\partial u_1}, \dots, \frac{\partial}{\partial u_{n-1}}\}$ is a basis of $G_{(\mathbf{u}, h(\mathbf{u}))}$. In this notation $(1, 0, \cdots, 0)$ is a coordinate representation in this basis for the vector field $\frac{\partial}{\partial u_1}$ for $(\mathbf{u}, h(\mathbf{u})) \in G$, yet its values at different $\mathbf{u}$ (or rather $(\mathbf{u}, h(\mathbf{u}))$) may not be identified as equal to each other. As a consequence, we may not have $D_{\mathbf{v}}(\frac{\partial}{\partial u_1}) = \mathbf{0}$ in contrast to the case if $\{\frac{\partial}{\partial u_i} = \mathbf{e}_i\}$ is used to represent a basis of the tangent space at a point in the ***flat*** Euclidean space.

Definition 7.3.4 Differential of a Function.

For any differentiable function f defined in $U \subset \mathbb{R}^n$ and a fixed $P \in U$, the operation

$$\mathbf{v} \mapsto D_{\mathbf{v}} f(P)$$

defines a linear function on $\mathbb{R}^n_P$, thus defines a **cotangent vector** in $\mathbb{R}^{n*}_P$, called the **differential** of f at P, and is denoted as $df(P)$. Thus

$$df(P)(\mathbf{v}) = D_{\mathbf{v}} f(P) \quad \text{for all } \mathbf{v} \in \mathbb{R}^n_P.$$

This definition requires f to be differentiable in a neighborhood of a point, it naturally defines a **field** of cotangent vectors in $\mathbb{R}^{n*}_P$ as P varies in this neighborhood. It is an example of a **one form**, and in this case, is called the differential of f.

When $f = x_i$ is a coordinate function, we find that

$$dx_i(\mathbf{v}) = D_{\mathbf{v}} x_i = v_i,$$

thus we find

$$df(P)(\mathbf{v}) = \sum_{i=1}^{n} v_i \frac{\partial f}{\partial x_i}(P) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(P) dx_i(\mathbf{v}) \quad \text{for all } \mathbf{v} \in \mathbb{R}^n_P,$$

and as one forms we have the classic formula

$$df = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} dx_i. \tag{7.3.1}$$

In older texts the differential df was often used interchangeably with $\sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \Delta x_i$ and referred to as the linear approximation to f at P. In the modern treatment, the differential df is a linear function on tangent vectors, so after taking a tangent vector $\mathbf{v}$ as input it gives the linear approximation $df(P)(\mathbf{v}) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} dx_i(\mathbf{v}) = \sum_{i=1}^{n} v_i \frac{\partial f}{\partial x_i}$ to f at P along $\mathbf{v}$.

Note that $\left\{ dx_1\Big|_P, \dots, dx_n\Big|_P \right\}$ forms the dual basis of $\left\{ \frac{\partial}{\partial x_1}\Big|_P, \dots, \frac{\partial}{\partial x_n}\Big|_P \right\}$, as

$$dx_i(\frac{\partial}{\partial x_j}) = D_{\frac{\partial}{\partial x_j}} x_i = \delta_{ij} \text{ for } 1 \le i, j \le n. \tag{7.3.2}$$

A general **one-form** can be expressed as $\sum_{i=1}^{n} \xi_i(\mathbf{x})dx_i$ for some scalar functions $\xi_i(\mathbf{x})$. It is called continuous (differentiable) if these functions are continuous (differentiable). We will learn later that for a general domain not all continuous one-forms can be the differential df of some continuously differentiable functions f.

The above discussion adapts easily to the context of a differentiable surface such as S represented in terms of a parametric representation $\vec{\gamma}(\mathbf{u})$ for $\mathbf{u} \in U \subset \mathbb{R}^k$. Then $\left\{ du_1\Big|_P, \ldots, du_k\Big|_P \right\}$ forms the dual basis of $\left\{ \frac{\partial}{\partial u_1}\Big|_P, \ldots, \frac{\partial}{\partial u_k}\Big|_P \right\}$. We will use du_i for $du_i\Big|_P$. Any one-form on S can be represented as $\sum_{i=1}^{k} \alpha_i(\mathbf{u})du_i$ for some functions $\alpha_i(\mathbf{u})$.

We next study the transformation laws of a vector field and a one-form under a change of coordinates. Suppose that $\vec{\gamma}^\dagger(\mathbf{v})$ for $\mathbf{v} = (v_1, \cdots, v_k) \in V \subset \mathbb{R}^k$ provides another parametrization for the same S via the relation $\mathbf{v} = F(\mathbf{u})$ for some continuously differentiable F with continuously differentiable inverse. A parametrization with this property is called **admissible**.

In this set tup, for any differentiable function f defined in a domain of $\mathbb{R}^n$ containing S, the chain rule implies

$$\frac{\partial(f \circ \vec{\gamma})(\mathbf{u})}{\partial u_j} = \sum_{i=1}^{k} \frac{\partial v_i}{\partial u_j} \frac{\partial(f \circ \vec{\gamma}^\dagger)(\mathbf{v})}{\partial v_i}.$$

$(f \circ \vec{\gamma})(\mathbf{u})$ and $(f \circ \vec{\gamma}^\dagger)(\mathbf{v})$ are simply two different coordinate representations of the same function f, so we have the following

$$\frac{\partial}{\partial u_j} = \sum_{i=1}^{k} \frac{\partial v_i}{\partial u_j} \frac{\partial}{\partial v_i}. \tag{7.3.3}$$

This is the transformation law between the two bases $\left\{ \frac{\partial}{\partial u_1}\Big|_P, \ldots, \frac{\partial}{\partial u_k}\Big|_P \right\}$ and $\left\{ \frac{\partial}{\partial v_1}\Big|_P, \ldots, \frac{\partial}{\partial v_k}\Big|_P \right\}$ for any $P \in \vec{\gamma}(U) \cap \vec{\gamma}^\dagger(V) \subset S$. Note that (7.3.3) is simply a form of the chain rule.

Suppose that X is a vector field on S, namely, $X(P) \in S_P$ is tangent to S at $P = \vec{\gamma}(\mathbf{u}) = \vec{\gamma}^\dagger(\mathbf{v})$,

$$X(P) = \sum_{i=1}^{k} a^i(\mathbf{u}) \frac{\partial}{\partial u_i} = \sum_{i=1}^{k} b^i(\mathbf{v}) \frac{\partial}{\partial v_i}, \tag{7.3.4}$$

then in addition to the relation

$$D_{X(P)} f = \sum_{i=1}^{k} b^i(\mathbf{v}) \frac{\partial(f \circ \vec{\gamma}^\dagger)(\mathbf{v})}{\partial v_i} = \sum_{j=1}^{k} a^j(\mathbf{u}) \frac{\partial(f \circ \vec{\gamma})(\mathbf{u})}{\partial u_j},$$

we also have

$$\begin{bmatrix} b^1(\mathbf{v}) \\ \vdots \\ b^k(\mathbf{v}) \end{bmatrix} = \left[\frac{\partial v_i}{\partial u_j} \right] \begin{bmatrix} a^1(\mathbf{u}) \\ \vdots \\ a^k(\mathbf{u}) \end{bmatrix}, \tag{7.3.5}$$

which follows from (7.3.4) and (7.3.3). (7.3.4) also encodes the geometric information that when $\mathbf{u}'(t) = (a^1(\mathbf{u}), \cdots, a^k(\mathbf{u}))$, then $\mathbf{v}(t) = F(\mathbf{u}(t))$ gives $\mathbf{v}'(t) = DF(\mathbf{u}(t))\mathbf{u}'(t)$, which is another form of (7.3.5).

Note that if $X(\mathbf{x})$ is a continuous vector field on S and $f(\mathbf{x})$ is a continuously differentiable function in a neighborhood of S. Then $D_{X(\mathbf{x})} f(\mathbf{x})$ is a continuous function on S, which can be computed via any admissible parametrization of S.

The dual of (7.3.3) is

$$dv_i = \sum_{j=1}^{k} \frac{\partial v_i}{\partial u_j}\, du_j. \tag{7.3.6}$$

Thus a one-form $\sum_{i=1}^{k} \xi_i(\mathbf{u}) du_i$ transforms to $\sum_{i=1}^{k} \eta_i(\mathbf{v}) dv_i$, where, using (7.3.6) we have

$$\sum_{i=1}^{k} \eta_i(\mathbf{v}) dv_i = \sum_{j=1}^{k}\sum_{i=1}^{k} \eta_i(\mathbf{v}) \frac{\partial v_i}{\partial u_j}\, du_j,$$

so it follows that

$$\xi_j(\mathbf{u}) = \sum_{i=1}^{k} \eta_i(\mathbf{v}) \frac{\partial v_i}{\partial u_j}. \tag{7.3.7}$$

In matrix notation, this transformation takes the form of

$$\begin{bmatrix} \xi_1(\mathbf{u}) \\ \vdots \\ \xi_k(\mathbf{u}) \end{bmatrix} = \left[\frac{\partial v_i}{\partial u_j}\right]^{\mathrm{T}} \begin{bmatrix} \eta_1(\mathbf{v}) \\ \vdots \\ \eta_k(\mathbf{v}) \end{bmatrix}.$$

Treating v_i as a function of $\mathbf{u}$, (7.3.6) is simply a case of (7.3.1). But we should read both sides of (7.3.6) as covectors in S_P^* with $P = \vec{\gamma}(\mathbf{u}) = \vec{\gamma}^\dagger(\mathbf{v})$.

As a consequence of (7.3.6), for any differentiable function f on S, its differential df computed in two different parametrization satisfy

$$df = \sum_{i=1}^{k} \frac{\partial f}{\partial u_i}\, du_i = \sum_{i=1}^{k} \frac{\partial f}{\partial v_i}\, dv_i.$$

Remark 7.3.5

In multivariable calculus $(\frac{\partial f}{\partial u_1}, \cdots, \frac{\partial f}{\partial u_k})$ *is called the gradient vector of* f *and points in the direction of greatest ascend of* f *with its magnitude* $\sqrt{\sum_{i=1}^{k} |\frac{\partial f}{\partial u_i}|^2}$ *representing the greatest ascend per unit length. Implicit in this statement is the use of Euclidean metric in the coordinate. If a change of coordinate is made, then the metric used to compute inner product between tangent vectors in the new coordinate may not take on the form of the usual Euclidean metric, and the transformed vector under (7.3.5) may not agree with* $(\frac{\partial f}{\partial v_1}, \cdots, \frac{\partial f}{\partial v_k})$.

The concept of a ***gradient vector*** *makes sense only with respect to a given metric. If* g *is a given metric then the gradient vector of* f *is defined to be* $\sharp(df)$ *with respect to the given metric* g*, namely,*

$$D_{\mathbf{v}} f = df(\mathbf{v}) = g(\sharp(df), \mathbf{v}).$$

In the context of Stokes Theorem, suppose that in the integrand $\xi(\mathbf{u}) \cdot \mathbf{u}'(t)$ of the line integral, we treat $\xi(\mathbf{u})$ as the coordinates of a one-form instead of a vector field, namely, $\Xi(\mathbf{u}) = \sum_{i=1}^{k} \xi_i(\mathbf{u})\, du_i$ and identify $\mathbf{u}'(t)$ with the tangent vector $\sum_{i=1}^{k} u_i'(t) \frac{\partial}{\partial u_i}$, then $\xi(\mathbf{u}) \cdot \mathbf{u}'(t) = \langle \Xi(\mathbf{u}), \mathbf{u}'(t) \rangle$ and under the change of variable $\mathbf{v} = F(\mathbf{u})$, $\Xi(\mathbf{u})$ transforms to $\sum_{i=1}^{k} \eta_i(\mathbf{v}) dv_i$ and $\mathbf{u}'(t)$ transforms to $\mathbf{v}'(t)$. We see that

$$\langle \Xi(\mathbf{u}), \mathbf{u}'(t) \rangle = \sum_{i=1}^{k} \eta_i(\mathbf{v}) v_i'(t).$$

This also follows directly from (7.3.5) and (7.3.7). Thus, when treated as a one-form, the line integral in Stokes Theorem is invariant under a change of variables.

Now that we have introduced tangent vectors and cotangent vectors, the algebraic tensor operations, including tensor product and exterior product, apply to them. Thus in addition to the tangent space S_P and cotangent space S_P^* at each point P of S, there are also spaces of higher order tensors. In an admissible parametrization $\mathbf{u} \in U \subset \mathbb{R}^k \mapsto \vec{\gamma}(\mathbf{u}) \in S$, $\{du_{i_1} \otimes \cdots \otimes du_{i_m} : 1 \le i_1, \cdots, i_m \le k\}$ forms a basis of the space $\mathcal{T}^m(S_{\vec{\gamma}(\mathbf{u})})$ of covariant tensors of order m of S at $\vec{\gamma}(\mathbf{u})$, while $\{du_{i_1} \wedge \cdots \wedge du_{i_m} : 1 \le i_1 < \cdots < i_m \le k\}$ forms a basis of the space $\Lambda^m(S_{\vec{\gamma}(\mathbf{u})})$ of covariant alternating tensors of order m of S at $\vec{\gamma}(\mathbf{u})$.

Example 7.3.6

In the case of two dimensions, if $u_1 = r, u_2 = \theta$ are the polar coordinates of $(x_1, x_2) \in \mathbb{R}^2$, then

$$\begin{aligned}\frac{\partial}{\partial r} &= \frac{\partial x}{\partial r}\frac{\partial}{\partial x} + \frac{\partial y}{\partial r}\frac{\partial}{\partial y} = \cos\theta\frac{\partial}{\partial x} + \sin\theta\frac{\partial}{\partial y},\\ \frac{\partial}{\partial \theta} &= \frac{\partial x}{\partial \theta}\frac{\partial}{\partial x} + \frac{\partial y}{\partial \theta}\frac{\partial}{\partial y} = -r\sin\theta\frac{\partial}{\partial x} + r\cos\theta\frac{\partial}{\partial y}.\end{aligned}$$

Noting that $\begin{bmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{bmatrix}$ being an orthogonal matrix, the above relation can be written as

$$\begin{bmatrix}\frac{\partial}{\partial r}\\ r^{-1}\frac{\partial}{\partial \theta}\end{bmatrix} = \begin{bmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{bmatrix}\begin{bmatrix}\frac{\partial}{\partial x}\\ \frac{\partial}{\partial y}\end{bmatrix},$$

from which we obtain

$$\begin{bmatrix}\frac{\partial}{\partial x}\\ \frac{\partial}{\partial y}\end{bmatrix} = \begin{bmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{bmatrix}\begin{bmatrix}\frac{\partial}{\partial r}\\ r^{-1}\frac{\partial}{\partial \theta}\end{bmatrix}.$$

Thus a vector field in rectangular coordinates $X(x,y)\frac{\partial}{\partial x} + Y(x,y)\frac{\partial}{\partial y}$, when represented in polar coordinates, becomes

$$\begin{aligned}&\begin{bmatrix}X(x,y) & Y(x,y)\end{bmatrix}\begin{bmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{bmatrix}\begin{bmatrix}\frac{\partial}{\partial r}\\ r^{-1}\frac{\partial}{\partial \theta}\end{bmatrix}\\ &= \left(X(r\cos\theta, r\sin\theta)\cos\theta + Y(r\cos\theta, r\sin\theta)\sin\theta\right)\frac{\partial}{\partial r}\\ &+r^{-1}\left(-X(r\cos\theta, r\sin\theta)\sin\theta + Y(r\cos\theta, r\sin\theta)\cos\theta\right)\frac{\partial}{\partial \theta}.\end{aligned}$$

If we treat $\mathbb{R}^2$ as equipped with the standard Euclidean metric, then $\{\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\}$ is orthonormal, so its dual basis $\{dx, dy\}$ is also orthonormal in the induced metric on cotangent space $\mathbb{R}^{2*}$. Since the relation between $\{\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\}$ and $\{\frac{\partial}{\partial r}, r^{-1}\frac{\partial}{\partial \theta}\}$ is via an orthogonal matrix, therefore the latter is also orthonormal in the tangent space. It follows that in metric notation we have $g(\frac{\partial}{\partial \theta}, \frac{\partial}{\partial \theta}) = r^2$.

Note that the dual basis of $\{\frac{\partial}{\partial r}, r^{-1}\frac{\partial}{\partial \theta}\}$ is $\{dr, rd\theta\}$. So $\{dr, rd\theta\}$ is orthonormal in the induced metric on cotangent space $\mathbb{R}^{2*}$. This can also be confirmed by the computation

$$dx \otimes dx + dy \otimes dy = dr \otimes dr + r^2 d\theta \otimes d\theta,$$

where one uses

$$dx = \cos\theta\, dr - r\sin\theta\, d\theta, \quad dy = \sin\theta\, dr + r\cos\theta\, d\theta. \tag{7.3.8}$$

Treating $r^2 d\theta \otimes d\theta$ as $(rd\theta)\otimes(rd\theta)$, one sees that $\{dr, rd\theta\}$ is an orthonormal basis. In metric notation we have $g(d\theta, d\theta) = r^{-2}$.

Note that if $\{\alpha_1, \cdots, \alpha_n\}$ and $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ are ***orthonormal dual*** basis, then $\sharp(\alpha_i) = \mathbf{u}_i$. It follows in our case that $\sharp(rd\theta) = r^{-1}\frac{\partial}{\partial\theta}$, so $\sharp(d\theta) = r^{-2}\frac{\partial}{\partial\theta}$.

We now treat the vector field $X(x,y)\frac{\partial}{\partial x} + Y(x,y)\frac{\partial}{\partial y}$ as $\sharp(X(x,y)\,dx + Y(x,y)\,dy)$, and the one form can be computed as

$$\begin{aligned} & X(x,y)\left(\frac{\partial x}{\partial r}\,dr + \frac{\partial x}{\partial\theta}\,d\theta\right) + Y(x,y)\left(\frac{\partial y}{\partial r}\,dr + \frac{\partial y}{\partial\theta}\,d\theta\right) \\ =& X(x,y)\ (\cos\theta\,dr - r\sin\theta\,d\theta) + Y(x,y)\ (\sin\theta\,dr + r\cos\theta\,d\theta) \\ =& \left(X(x,y)\cos\theta + Y(x,y)\sin\theta\right)\ dr + r\left(-X(x,y)\sin\theta + Y(x,y)\cos\theta\right)\ d\theta, \end{aligned}$$

from which we can apply the $\sharp$ operation to get the same result. In addition, if Γ is a parametric curve in $\mathbb{R}^2$, then the line integral

$$\begin{aligned} & \int_\Gamma \{X(x,y)\,dx + Y(x,y)\,dy\} \\ =& \int_\Gamma \{(X(x,y)\cos\theta + Y(x,y)\sin\theta)\ dr + r\left(-X(x,y)\sin\theta + Y(x,y)\cos\theta\right)\ d\theta\}\,. \end{aligned}$$

Note that (7.3.8) also gives

$$dx \wedge dy = (\cos\theta\,dr - r\sin\theta\,d\theta)\wedge(\sin\theta\,dr + r\cos\theta\,d\theta) = r\,dr\wedge d\theta.$$

Finally, for a differentiable function f,

$$df = \frac{\partial f}{\partial x}\,dx + \frac{\partial f}{\partial y}\,dy = \frac{\partial f}{\partial r}\,dr + \frac{\partial f}{\partial\theta}\,d\theta,$$

so taking $\sharp$ operation gives

$$\frac{\partial f}{\partial x}\frac{\partial}{\partial x} + \frac{\partial f}{\partial y}\frac{\partial}{\partial y} = \frac{\partial f}{\partial r}\frac{\partial}{\partial r} + r^{-2}\frac{\partial f}{\partial\theta}\frac{\partial}{\partial\theta}.$$

This gives the gradient of f in the polar coordinate as $(\frac{\partial f}{\partial r}, r^{-2}\frac{\partial f}{\partial\theta})$.

One also notes that $\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2 \neq \left(\frac{\partial f}{\partial r}\right)^2 + \left(\frac{\partial f}{\partial\theta}\right)^2$ in general. Instead,

$$\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2 = \left(\frac{\partial f}{\partial r}\right)^2 + \frac{1}{r^2}\left(\frac{\partial f}{\partial\theta}\right)^2.$$

Exercises

1. Using the relation between rectangular and spherical polar coordinates in $\mathbb{R}^3$:

$$\begin{cases} x_1 &= r\sin\theta\cos\phi \\ x_2 &= r\sin\theta\sin\phi \\ x_3 &= r\cos\theta \end{cases}$$

to determine dx_1, dx_2, dx_3 in terms of $dr, d\theta, d\phi$. Then for a differentiable function f determine $\frac{\partial f}{\partial r}, \frac{\partial f}{\partial\theta}, \frac{\partial f}{\partial\phi}$ in terms of $\frac{\partial f}{\partial x_i}, i = 1, 2, 3$.

Hint. Use $df = \sum_{i=1}^{3} \frac{\partial f}{\partial x_i} dx_i$ and substitute dx_i in terms of $dr, d\theta, d\phi$ to identify $\frac{\partial f}{\partial r}, \frac{\partial f}{\partial \theta}, \frac{\partial f}{\partial \phi}$.

2. On the sphere $\sum_{i=1}^{3} x_i^2 = 1$, consider (x_1, x_2) and (θ, ϕ) as coordinates for a portion of the upper hemisphere. Find the relations between dx_1, dx_2 and $d\theta, d\phi$, as well as between $\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}$ and $\frac{\partial}{\partial \theta}, \frac{\partial}{\partial \phi}$. Identify a choice of the domain on which the change of coordinates between these two sets of coordinates is continuously differentiable and has continuously differentiable inverse.

3. Rewrite the system of ODE

$$\begin{cases} x'(t) & = -\frac{\partial f(x,y)}{\partial x} \\ y'(t) & = -\frac{\partial f(x,y)}{\partial y} \end{cases}$$

as a system in the polar coordinate (r, θ) of (x, y) and the partial derivatives of f with respect to r and θ.

7.3.2 Tangent Map/Differential and its Pull-Back of a Differentiable Map

Suppose that $F : U \subset \mathbb{R}^m \mapsto \mathbb{R}^n$ is differentiable, then for any $P \in U$, $DF(P)$ is a linear map from $\mathbb{R}^m_P$ to $\mathbb{R}^n_{F(P)}$ mapping $\mathbf{v}$ to $DF(P)\mathbf{v}$, and is denoted as $F_*(P)$, and is often called the **tangent map** (or even called the differential and denoted as dF) of F at P.

The action of $F_*(P)$ can also be seen through how a differentiable function f on $\mathbb{R}^n$ is differentiated through F:

$$D_{\frac{\partial}{\partial u_i}} (f \circ F) = D_{F_*\left(\frac{\partial}{\partial u_i}\right)} f,$$

namely, to treat $F_*\left(\frac{\partial}{\partial u_i}\right)$ as tangent vector $D_{u_i}F(P)$ at $F(P)$. When F is a parametrization map, we have identified in our notation $F_*\left(\frac{\partial}{\partial u_i}\right)$ with $\left(\frac{\partial}{\partial u_i}\right)$.

Note that $F_*(P)$ is determined in terms of the first derivatives of F, so a more accurate notation would have been $DF(P)$ or $dF(P)$, but it has been a traditional to use $DF(P)$ as the matrix representation of F_* under the chosen coordinates.

This notion and notation turn out to be very useful. Suppose $F(P) = Q$. Write out F in components

$$x_i = F_i(\mathbf{u}) = F_i(u_1, \ldots, u_m), 1 \le i \le n.$$

Then each F_i is a differentiable function of $\mathbf{u}$, thus

$$dF_i(\mathbf{u}) = \sum_{j=1}^{m} \frac{\partial F_i}{\partial u_j} du_j.$$

The geometric interpretation of the linear map $F_* : \mathbb{R}^m_P \mapsto \mathbb{R}^n_Q$ is seen as follows. For any $\mathbf{v} \in \mathbb{R}^m_P$, $P + t\mathbf{v}$ is a curve in $\mathbb{R}^m$ passing through P with tangent $\mathbf{v}$ at $t = 0$, and $\mathbf{x}(t) := F(P + t\mathbf{v})$ is a curve in $\mathbb{R}^n$ passing through Q at $t = 0$, its tangent at $t = 0$ is

$$\mathbf{x}'(0) = DF(P)\mathbf{v}.$$

Thus $\mathbf{x}'(0) = F_*(\mathbf{u})$. In components we see

$$x_i'(0) = \sum_{j=1}^{m} v_j \frac{\partial F_i}{\partial u_j}(P) = dF_i(P)(\mathbf{v}).$$

This is a reason for using dF as a common notation for F_*. Another way of writing this relation is to note that

$$\begin{aligned} d(f\circ F)(\mathbf{v}) =& D_{\mathbf{v}}(f\circ F) \\ =& \sum_{j=1}^{m} v_j \frac{\partial(f\circ F)}{\partial u_j} \\ =& \sum_{j=1}^{m} v_j \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\frac{\partial F_i}{\partial u_j} \\ =& \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \sum_{j=1}^{m} v_j \frac{\partial F_i}{\partial u_j} \\ =& \sum_{i=1}^{n} x_i'(0)\frac{\partial f}{\partial x_i} \\ =& D_{\mathbf{w}} f \\ =& df(\mathbf{w}) \\ =& df(dF(\mathbf{v})), \end{aligned}$$

where $\mathbf{w} = dF(\mathbf{v})$. Namely,

$$d(f\circ F) = df \circ dF \text{ and } D_{\mathbf{v}}(f\circ F) = D_{dF(\mathbf{v})} f.$$

It is easier to understand this relation in terms of the following diagram.

$$d(f\circ F) : \mathbb{R}^m_P \xmapsto{dF} \mathbb{R}^n_Q \xmapsto{df} \mathbb{R}_{f(Q)}.$$

Using the dual maps (or pull-backs), we have

$$\mathbb{R}^*_{f(Q)} \xmapsto{(df)^*} \mathbb{R}^{n*}_Q \xmapsto{(dF)^*} \mathbb{R}^{m*}_P,$$

where, if we denote $z = f(\mathbf{x})$, then $(df)^*(dz) = \sum_{i=1}^n \frac{\partial f}{\partial x_i} dx_i$, and

$$(dF)^*(dx_i) = \sum_{j=1}^{m} \frac{\partial x_i}{\partial u_j} du_j.$$

We note that many books use f^* to denote $(df)^*$, and F^* to denote $(dF)^*$. Thus we have the corresponding relation $(f\circ F)^* = F^* \circ f^*$, and the pull-back operation behaves like substitution in the differential: $f^*(dz) = df(\mathbf{x}) = \sum_{i=1}^n \frac{\partial f}{\partial x_i} dx_i$, etc.

Remark 7.3.7

Some of these definitions and operational rules may seem abstract and formal. But they are mostly the chain rule in different disguise and capture the essence of differential calculus. In particular, the pull-back operation is essentially the substitution rule for differentials: $dx_i, i = 1, \cdots, n$ *may be considered as differential forms in* $\mathbb{R}^n$*, and when* $x_i = F_i(\mathbf{u})$ *is a differentiable function of* $\mathbf{u} \in \mathbb{R}^m$, $(dF_i)^*(dx_i)$ *is just to substitute* $x_i = F_i(\mathbf{u})$ *to get* $dx_i = \sum_{j=1}^m \frac{\partial F_i}{\partial u_j} du_j$ *as a one form in the* $\mathbf{u}$ *space.*

Part of the reason for the complication of the notation is that it needs to reflect the dependence on the base point and the tangent vector. For instance, $(dF(\mathbf{u}))^*(\alpha)$ *is well defined when* $\alpha \in \mathbb{R}^{n*}_{F(\mathbf{u})}$ *and needs to act on a tangent vector* $\mathbf{v} \in \mathbb{R}^m_{\mathbf{u}}$.

Relying entirely on parentheses as delimiters for different inputs and operations can be tiring, so we often use braket notation $\langle \alpha, \mathbf{v} \rangle$ *to denote* $\alpha(\mathbf{v})$*. In this notation, we have*

$$\langle (dF(\mathbf{u}))^* (\alpha), \mathbf{v} \rangle = \langle \alpha, dF(\mathbf{u})(\mathbf{v}) \rangle.$$

The pull-back operation $(dF(\mathbf{u}))^*$ *has the advantage that for any differential form* $\alpha(\mathbf{x})$ *defined on* $\mathbb{R}^n$ *(or on a surface in* $\mathbb{R}^n$*),* $(dF(\mathbf{u}))^* (\alpha)$ *is a well defined a differential form in the* $\mathbf{u}$ *variable in* $\mathbb{R}^m$ *making it possible for us to do computations in the Euclidean parameter domain, while for a vector field* $V(\mathbf{u})$ *in the* $\mathbf{u}$ *variable,* $dF(\mathbf{u})(V(\mathbf{u}))$ *defines a vector at* $\mathbb{R}^n_{F(\mathbf{u})}$*, but it may not be considered as a vector field in the* $\mathbf{x}$ *variable, as there may be* $\mathbf{u}' \neq \mathbf{u}$ *such that* $F(\mathbf{u}') = F(\mathbf{u})$*, and* $dF(\mathbf{u}')(V(\mathbf{u}'))$ *may not equal* $dF(\mathbf{u})(V(\mathbf{u}))$*.*

Remark 7.3.8

Our consideration for the dual space of the space of tangent vectors and the pull-back maps on tensors is because they provide a natural setting for us to relate M *to* N *whenever there is a differentiable map* $F : M \mapsto N$*. For our specific problem of developing an integration theory on differential forms, we often take* M *to be a standard cell (such as a cube or simplex) in* $\mathbb{R}^m$*, the pull-back maps allow us to represent forms on* N *as forms on the cell in the Euclidean space and use the developed integration theory in the Euclidean space as a tool.*

Exercises

1. Consider $(x_1, x_2, x_3) = F(r, \theta, \phi)$ given by

$$\begin{cases} x_1 &= r \sin\theta \cos\phi \\ x_2 &= r \sin\theta \sin\phi \\ x_3 &= r \cos\theta. \end{cases}$$

(a) Compute $F_*(\frac{\partial}{\partial r}), F_*(\frac{\partial}{\partial \theta}), F_*(\frac{\partial}{\partial \phi})$.

(b) Compute $F^*(dx_i)$, $F^*(dx_1 \wedge dx_2)$, $F^*(dx_2 \wedge dx_3)$ and $F^*(dx_3 \wedge dx_1)$.

(c) Compute $F^*(dx_1 \otimes dx_i + dx_2 \otimes dx_2 + dx_3 \otimes dx_3)$.

(d) Let $G(\theta, \phi) = F(1, \theta, \phi)$. Compute $G_*(\frac{\partial}{\partial \theta}), G_*(\frac{\partial}{\partial \phi})$, $G^*(dx_i)$, $G^*(dx_1 \wedge dx_2)$, $G^*(dx_2 \wedge dx_3)$, and $G^*(dx_3 \wedge dx_1)$.

(e) Compute $G^*(x_1 dx_1 + x_2 dx_2 + x_3 dx_3)$ and $G^*(dx_1 \otimes dx_i + dx_2 \otimes dx_2 + dx_3 \otimes dx_3)$.

2. Let $(x, y, z) = F(u, v) = (u^2 - v^2, 2uv, 1)$. Compute $F_*(\frac{\partial}{\partial u})$, $F^*(ydx + zdy)$ and $F^*(dx \wedge dy + dy \wedge dz)$.

7.4 Exterior Differential Operator d and Boundary Operator ∂

We now introduce the two most central objects in the study of integration of differential forms, the exterior differential operator d on differential forms and boundary operator ∂ on **singular cubes** or **chains**.

7.4.1 Exterior Differential Operator

We use $\mathbb{R}^m_P$ to denote the tangent space at P. Let $T\mathbb{R}^m = \{(P, \mathbf{v}) : P \in \mathbb{R}^m, \mathbf{v} \in \mathbb{R}^m_P\}$ denote the **tangent bundle** of $\mathbb{R}^m$, and $\Lambda^k(T\mathbb{R}^m) = \{(P, \omega) : P \in \mathbb{R}^m, \omega \in \Lambda^k(\mathbb{R}^m_P)\}$ denote the set of k-forms on $\mathbb{R}^m$: it assigns an alternating tensor of order k at each point of $\mathbb{R}^m$. Such a form is called differentiable if the coefficients of the form in the standard basis $\{dx_{i_1} \wedge \cdots \wedge dx_{i_k}\}$ are differentiable functions. In fact, we will take $\Lambda^k(T\mathbb{R}^m)$ to mean the set of k-forms that are infinitely times differentiable. The slight complication of this notation is to give indication that elements in $T\mathbb{R}^m$ and in $\Lambda^k(T\mathbb{R}^m)$ have a value at each base point, while the notation $\Lambda^k(\mathbb{R}^m)$ does not have any relation with a base point.

For a differentiable function f, $df = \sum_{i=1}^m \frac{\partial f}{\partial x_i}\, dx_i$ may be regarded as the output of a linear operator

$$d : f \in \Lambda^0(T\mathbb{R}^m) := C^\infty(\mathbb{R}^m) \mapsto df \in \Lambda^1(T\mathbb{R}^m).$$

We now extend this operator to

$$d : \omega \in \Lambda^k(T\mathbb{R}^m) \mapsto d\omega \in \Lambda^{k+1}(T\mathbb{R}^m)$$

for $1 \leq k \leq m$. Any $\omega \in \Lambda^1(T\mathbb{R}^m)$ can be expressed as $\omega(\mathbf{x}) = \sum_{i=1}^m \omega_i(\mathbf{x}) dx_i$. We define

$$d\omega(\mathbf{x}) = \sum_{i=1}^m d\omega_i(\mathbf{x}) \wedge dx_i.$$

Since $d\omega_i(\mathbf{x}) = \sum_{j=1}^m \frac{\partial \omega_i}{\partial x_j} dx_j$, we have

$$d\omega(\mathbf{x}) = \sum_{j<i} \left(\frac{\partial \omega_i}{\partial x_j} - \frac{\partial \omega_j}{\partial x_i} \right) dx_j \wedge dx_i.$$

Recall that our discussion on the curl of a vector field in higher dimensions leads us to this object: if we treat the vector field $X(\mathbf{x})$ as a one form $\omega(\mathbf{x})$, then the curl of $X(\mathbf{x})$ corresponds to $d\omega(\mathbf{x})$.

For $k \geq 2$, it turns out to be natural to define d in a similar way: express

$$\omega(\mathbf{x}) = \sum_{1 \leq i_1 < i_2 < \ldots < i_k \leq m} \omega_{i_1 i_2 \ldots i_k}(\mathbf{x})\ dx_{i_1} \wedge dx_{i_2} \wedge \ldots \wedge dx_{i_k}$$

and define

$$d\omega(\mathbf{x}) = \sum_{1 \leq i_1 < i_2 < \ldots < i_k \leq m} d\omega_{i_1 i_2 \ldots i_k}(\mathbf{x}) \wedge dx_{i_1} \wedge dx_{i_2} \wedge \ldots \wedge dx_{i_k}.$$

In other words, for any single term $\omega_{i_1 i_2 \ldots i_k}(\mathbf{x}) dx_{i_1} \wedge dx_{i_2} \wedge \ldots \wedge dx_{i_k}$, its exterior differential is simply

$$d\omega_{i_1 i_2 \ldots i_k}(\mathbf{x}) \wedge dx_{i_1} \wedge dx_{i_2} \wedge \ldots \wedge dx_{i_k}.$$

We will discuss soon why this definition is natural.

For example, for such an ω with $k = 2$ in $\mathbb{R}^3$, using

$$\begin{aligned}
&d\left(\omega_{12}(\mathbf{x})dx_1 \wedge dx_2\right) \\
=&\left(\frac{\partial\omega_{12}}{\partial x_1}\,dx_1 + \frac{\partial\omega_{12}}{\partial x_2}\,dx_2 + \frac{\partial\omega_{12}}{\partial x_3}\,dx_3\right) dx_1 \wedge dx_2 \\
=&\frac{\partial\omega_{12}}{\partial x_3} dx_3 \wedge dx_1 \wedge dx_2 \\
=&\frac{\partial\omega_{12}}{\partial x_3} dx_1 \wedge dx_2 \wedge dx_3
\end{aligned}$$

and similar computations we get

$$\begin{aligned}
&d\left(\omega_{12}(\mathbf{x})dx_1 \wedge dx_2 + \omega_{23}(\mathbf{x})dx_2 \wedge dx_3 + \omega_{13}(\mathbf{x})dx_1 \wedge dx_3\right) \\
=&\frac{\partial\omega_{12}}{\partial x_3} dx_3 \wedge dx_1 \wedge dx_2 + \frac{\partial\omega_{23}}{\partial x_1} dx_1 \wedge dx_2 \wedge dx_3 + \frac{\partial\omega_{13}}{\partial x_2} dx_2 \wedge dx_1 \wedge dx_3 \\
=&\left(\frac{\partial\omega_{12}}{\partial x_3} + \frac{\partial\omega_{23}}{\partial x_1} - \frac{\partial\omega_{13}}{\partial x_2}\right) dx_1 \wedge dx_2 \wedge dx_3.
\end{aligned}$$

The most important properties of this operator are summarized in the following theorem; the central ones are the last two.

Theorem 7.4.1 Properties of the Exterior Differential Operater d.

1. *$d(\omega + \eta) = d\omega + d\eta$ for any $\omega, \eta \in \Lambda^k(T\mathbb{R}^m)$.*
2. *$d(\omega \wedge \eta) = d\omega \wedge \eta + (-1)^k \omega \wedge d\eta$ for any $\omega \in \Lambda^k(T\mathbb{R}^m), \eta \in \Lambda^l(T\mathbb{R}^m)$.*
3. *$d \circ d = 0$, namely, $d(d\omega) = 0$ for any $\omega \in \Lambda^k(T\mathbb{R}^m)$.*
4. *$F^*(d\omega) = d(F^*(\omega))$ for any differentiable map F.*

In the case of $\omega(\mathbf{x}) = \sum_{i=1}^m \omega_i(\mathbf{x})dx_i$, the property $d \circ d = 0$ is demonstrated as

$$\begin{aligned}
&d\left(\sum_{i,\,j=1}^n \frac{\partial\omega_i}{\partial x_j} dx_j \wedge dx_i\right) \\
=&\sum_{i,\,j=1}^n d\left(\frac{\partial\omega_i}{\partial x_j}\right) \wedge dx_j \wedge dx_i \\
=&\sum_{i=1}^n \sum_{j,\,k=1}^n \left(\frac{\partial^2\omega_i}{\partial x_j\,\partial x_k}\right) dx_k \wedge dx_j \wedge dx_i \\
=&\sum_{i=1}^n \sum_{k<j}^n \left(\frac{\partial^2\omega_i}{\partial x_j\,\partial x_k} - \frac{\partial^2\omega_i}{\partial x_k\,\partial x_j}\right) dx_k \wedge dx_j \wedge dx_i = 0
\end{aligned}$$

This can also be seen by applying the rules of exterior differentiation:

$$\begin{aligned}
d(d\omega) =&d(\sum_{i=1}^n d\omega_i(\mathbf{x}) \wedge dx_i) \\
=&\sum_{i=1}^n \left(d\left(d\omega_i(\mathbf{x})\right) \wedge dx_i - d\omega_i(\mathbf{x}) \wedge d(dx_i)\right) = 0
\end{aligned}$$

using $d\left(d\omega_i(\mathbf{x})\right) = 0$ and $d(dx_i) = 0$, which can be verified more easily.

Exercises

1. Let $\omega(x, y, z) = P(x, y, z)\, dy \wedge dz$ be a continuous 2-form in $\mathbb{R}^3$ and $F(u, v) = (f_1(u, v), f_2(u, v), f_3(u, v))$ be a differentiable map from $\mathbb{R}^2$ to $\mathbb{R}^3$.

(a) Compute $d\omega$, $d(d\omega)$, $F^*(\omega)$, $F^*(d\omega)$ and $dF^*(\omega)$.

(b) If $f_3(u, v) =$ const. for $(u, v) \in \mathbb{R}^2$, verify that $F^*(\omega) = 0$ and $F^*(d\omega) = 0$.

2. $F(u, v, w) = (f_1(u, v, w), f_2(u, v, w), f_3(u, v, w))$ be a differentiable map from $\mathbb{R}^3$ to $\mathbb{R}^3$. Compute $F^* \left(G(x, y, z)dx \wedge dy \wedge dz\right)$.

3. Compute $d\left[(P\, dx + Q\, dy + R\, dz) \wedge dz\right]$.

7.4.2 The Boundary Operator, Singular Chain, and Stokes Theorem

Since k-forms act on k-tuples of vectors, or k-dimensional subspaces, we now discuss the geometric objects which interact with k-forms. Our ultimate goal is to introduce k-dimensional surfaces or manifolds. A proper definition of the latter requires some preparation. Informally we can think of a k-dimensional surface as pieced together by reasonably well behaved patches. Here "reasonably well behaved" means that it can be represented by a parametric map defined on a simple domain in $\mathbb{R}^k$, and a simple domain is meant either the standard cube $I^k := [0, 1] \times \cdots \times [0, 1]$ or the standard simplex Q^k (another commonly used notation is Δ^k), defined as $\{(x_1, \ldots, x_k) : x_i \geq 0 \text{ for each } i = 1, \ldots, k, \text{ and } \sum_{i=1}^{k} x_i \leq 1\}$. The latter is a generalization of triangle in $\mathbb{R}^2$ and tetrahedron in $\mathbb{R}^3$.

Piecing together the patches would require consideration of the "faces" and other lower dimensional edges of the patches. Both the cube and the simplex have relative simple description of their faces and lower dimensional edges. For the purpose of developing integration, the cube is easier, for the integration limits are easy. But it is even easier to catalog all lower dimensional faces and edges of the simplex, for the simplex Q^k has $k + 1$ vertices defined by requiring any k of the $k + 1$ inequalities in the definition of Q^k to be equalities, and any l dimensional edge/face of Q^k corresponds to setting $k - l$ of the $k + 1$ inequalities in the definition of Q^k to be equalities, which has $l + 1$ vertices. For I^k, its $k - 1$ dimensional faces are easy to describe: it has one pair corresponding to each x_i being 0 or 1; however, it has 2^k vertices vs the $k + 1$ vertices of Q^k, and not any choice of $l + 1$ vertices of I^k correspond to an l dimensional edge/face of I^k.

We will follow Spivak to focus on using I^k as our standard domain. An actual theory of piecing together patches modeled on either I^k or Q^k is difficult. In two dimension, it amounts to showing that any surface (to be properly defined) can be triangulated. One way to bypass this task is to use a **partition of unity** to write the integrand—a differential form in this context—as the sum of some terms, each of which has support in a cube.

For now we will only focus on how a k-form ω defined on $A \subset \mathbb{R}^n$ interacts with a single k-dimensional patch in A, called a **singular k-cube**, defined as a continuous (or differentiable) map $c : I^k \mapsto A \subset \mathbb{R}^n$. We will not require c to be bijective, or the Jacobian of c to have rank k everywhere, so $c(I^k)$ may not be a k-dimensional geometric object, or has a clear lower dimensional faces/edges. But we will use c to pull back ω to I^k, then $c^*(\omega)$ becomes a k-form on I^k, and I^k has clearly defined lower dimensional faces/edges.

In fact, since our set up was to study the integration of a k-form ω on the k-dimensional boundary faces of a singular $(k + 1)$-cube, we will change c to be a singular $(k + 1)$-cube now. I^{k+1} has $2(k + 1)$ k-dimensional faces, and each face of I^{k+1} has identical geometry as I^k. Since $x_i = 0$ or 1 gives rise to a face of c, we denote such a face by $c_{i,0}$ or $c_{i,1}$ respectively. Technically define

$$F_{i,0}(x_1, \ldots, x_k) = (x_1, \ldots, x_{i-1}, 0, x_i, \ldots, x_k),$$

namely, setting the ith component of $F_{i,0}(x_1,\dots,x_k)$ to be 0, and assigning the components in the $i+1$ through $k+1$ position as $(x_i,\dots,x_k)$. Similarly for $F_{i,1}$. Then $c_{i,0}=c\circ F_{i,0}$ and $c_{i,1}=c\circ F_{i,1}$ each defines a singular k-cube in A.

We will let s to take either 0 or 1 and integrate $c^*_{i,s}(\omega)=f_{i,s}(\mathbf{x})dx_1\wedge\cdots\wedge dx_k$ on I^k and define

$$\int_{c_{i,s}}\omega := \int_{I^k} c^*_{i,s}(\omega)=\int_{I^k} f_{i,s}(\mathbf{x})dx_1\cdots dx_k.$$

To consider the effect of the integration of ω on all faces of I^{k+1} we will consider an appropriate algebraic sum of these integrals. The result amounts to considering some algebraic sums of the $c_{i,s}$'s, and leads to the definition of a k-**chain** as a finite sum of singular k-cubes with integer coefficients.

Definition 7.4.2 The Boundary of a Singular Cube and a Singular Chain.

The **boundary** of the singular $(k{+}1)$-cube c is defined to be $\partial c=\sum_{i=1}^{k+1}\sum_{s=0}^{1}(-1)^{i+s}c_{i,s}$.

The boundary of a k-chain, $m_1c_1+\ldots+m_rc_r$, is defined as $m_1\partial c_1+\ldots+m_r\partial c_r$.

The reason for this choice of sign will become clear in the proof of Stokes Theorem.

Remark 7.4.3

The sum in the definition of the boundary is a formal sum of maps from I^k into to A, not as a sum of vector-valued functions. For one thing, if $c_1,c_2: I^k\mapsto A$, then as the sum of vector-valued functions c_1+c_2 may not take values in A. Secondly, we are going to use c_1+c_2 only in the context of how they pull back forms on A to those on I^k: $(c_1+c_2)^*(\omega)=c_1^*(\omega)+c_2^*(\omega)$.

Let's take the case of $c(x,y)=(x,y^2)$ for $(x,y)\in I^2$, $A=I^2$ and $c_{1,0}(t)=c(0,t)=(0,t^2)$, $c_{1,1}(t)=c(1,t)=(1,t^2)$. Then the vector-valued algebraic sum $c_{1,1}(t)-c_{1,0}(t)$ of $c_{1,0}(t)$ and $c_{1,1}(t)$ is the map $c(t)=(1,0)$. Then for any differential form ω defined in $(x,y)\in I^2$, $c^*(\omega)=0$, while if we take $\omega=x\,dy$, then $(c_{1,1}-c_{1,0})^*(x\,dy)=1\,d(t^2)-0\,d(t^2)=2t\,dt$. If we consider the vector-valued sum $c_{1,1}(t)+c_{1,0}(t)$, it is $t\mapsto(1,2t^2)$, which may not take value in $A=I^2$ for a certain range of t.

As a simple illustration of the integral of a k-form on a k-cube or chain, consider the case of $k=1$: suppose that $\omega(\mathbf{x})=\sum_{i=1}^n\omega_i(\mathbf{x})dx_i$ is a 1-form in $\mathbb{R}^n$, and $c:[0,1]\mapsto\mathbb{R}^n$ is a 1-cube, then $c^*(\omega)=\sum_{i=1}^n\omega_i(c(t))c_i'(t)dt$, so

$$\int_c\omega=\int_{[0,1]}\sum_{i=1}^n\omega_i(c(t))c_i'(t)dt$$

In the case that $\omega=df$ for some differentiable function $f(\mathbf{x})$, $\omega_i(\mathbf{x})=D_if(\mathbf{x})$, and $\sum_{i=1}^n D_if(c(t))c_i'(t)=[f(c(t))]'$, so

$$\int_{[0,1]}\sum_{i=1}^n\omega_i(c(t))c_i'(t)dt=\int_{[0,1]}[f(c(t))]'\,dt=f(c(1))-f(c(0))=\int_{\partial c}f.$$

Once these definitions are properly developed, we are be ready to prove a version of Stokes Theorem on a singular cube (or chain).

Theorem 7.4.4 Stokes Theorem on Singular Chains.

If ω is a k-form in A, and c is a singular $(k+1)$-cube (or chain), then

$$\int_{\partial c} \omega = \int_c d\omega.$$

Proof. $c^*(\omega)$ is a k-form on I^{k+1}, so can be written as

$$c^*(\omega)(\mathbf{x}) = \sum_{i=1}^{k+1} f_i(\mathbf{x}) dx_1 \wedge \cdots \wedge \widehat{dx_i} \wedge \cdots \wedge dx_{k+1},$$

where $\widehat{dx_i}$ means that the dx_i is omitted in the wedge product. Then

$$c^*(d\omega)(\mathbf{x}) = d\left[c^*(\omega)\right](\mathbf{x}) = \sum_{i=1}^{k+1} df_i(\mathbf{x}) \wedge dx_1 \wedge \cdots \widehat{dx_i} \wedge \cdots \wedge dx_{k+1}.$$

We claim that

$$c^*_{i,0}(\omega) = f_i(x_1, \ldots, x_{i-1}, 0, x_i, \ldots, x_k) dx_1 \wedge \cdots \wedge dx_k,$$

and

$$c^*_{i,1}(\omega) = f_i(x_1, \ldots, x_{i-1}, 1, x_i, \ldots, x_k) dx_1 \wedge \cdots \wedge dx_k.$$

This is because $F^*_{i,0}(dx_i) = 0$ and $F^*_{i,1}(dx_i) = 0$ (Heuristically, on these faces x_i is a constant, so $dx_i = 0$ when applied to tangent vectors in these faces), so, for $j \neq i$,

$$F^*_{i,0}(dx_1 \wedge \cdots \widehat{dx_j} \wedge \cdots \wedge dx_{k+1}) = 0,$$

and $c^*_{i,0}(\omega)$ will have only one term arising from $f_i(\mathbf{x}) dx_1 \wedge \cdots \wedge \widehat{dx_i} \wedge \cdots \wedge dx_{k+1}$ with $x_i = 0$ in $f_i(\mathbf{x})$:

$$c^*_{i,0}(\omega) = F^*_{i,0}(c^*(\omega)) = f_i(x_1, \ldots, x_{i-1}, 0, x_i, \ldots, x_k) dx_1 \wedge \cdots \wedge dx_k.$$

The same argument works for $c^*_{i,1}(\omega)$.

Note that

$$df_i(\mathbf{x}) \wedge dx_1 \wedge \cdots \wedge \widehat{dx_i} \wedge \cdots \wedge dx_{k+1} = (-1)^{i+1} D_i f_i(\mathbf{x}) dx_1 \wedge \cdots \wedge dx_i \wedge \cdots \wedge dx_{k+1},$$

so

$$\begin{aligned} &\int_{I^{k+1}} df_i(\mathbf{x}) \wedge dx_1 \wedge \cdots \widehat{dx_i} \wedge \cdots \wedge dx_{k+1} \\ =&(-1)^{i+1} \int_{I^{k+1}} D_i f_i(\mathbf{x}) dx_1 \wedge \cdots dx_i \wedge \cdots \wedge dx_{k+1} \end{aligned}$$

But integrating out the i-th coordinate first gives

$$\begin{aligned} &\int_{I^{k+1}} D_i f_i(\mathbf{x}) dx_1 \wedge \cdots dx_i \wedge \cdots \wedge dx_{k+1} \\ =&\int_{I^k} f_i(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_{k+1}) \Big|_{x_i=0}^{x_i=1} dx_1 \cdots \widehat{dx_i} \cdots dx_{k+1} \end{aligned}$$

so

$$\int_{I^{k+1}} df_i(\mathbf{x}) \wedge dx_1 \wedge \cdots \widehat{dx_i} \wedge \cdots \wedge dx_{k+1}$$

$$
\begin{aligned}
=&(-1)^{i+1}\int_{I^k}\left(F_{i,1}^*(c^*(\omega)) - F_{i,0}^*(c^*(\omega))\right)\\
=&(-1)^{i+1}\int_{I^k}\left(c_{i,1}^*(\omega) - c_{i,0}^*(\omega)\right)
\end{aligned}
$$

It is now clear from the definition of ∂c that

$$\int_{\partial c}\omega = \sum_{i=1}^{k+1}\sum_{s=0}^{1}(-1)^{i+s}\int_{I^k} c_{i,s}^*(\omega) = \int_{I^{k+1}} dc^*(\omega) = \int_c d\omega.$$

■

Remark 7.4.5

This version of the Stokes Theorem is formulated as integrals on standard cells I^k and I^{k+1} of forms pulled back by a map c defined on I^{k+1}. In concrete situations c often defines a $(k+1)$-dimensional differentiable surface with the restriction of c to the boundary of I^{k+1} defining k-dimensional differentiable surfaces. In such cases the integrals can be considered as defined on these geometric surfaces. The $k = 1$ case gives the classical Stokes Theorem for a surface parametrized via a map from I^2– see the example below for details. If we encounter more complicated surfaces in applications, we need to partition the surface as the non-overlapping union of several pieces which may share some common edges and each piece can be parametrized via a map from I^2, one then needs to account for the contributions from the difference pieces in the boundary integral $\int_{\partial c}\omega$.

In some sense, the operator d and the boundary ∂c are motivated and defined to make the Stokes Theorem hold in the general setting.

The ∂ operator also has the property that $\partial\circ\partial = 0$. It is a reflection of the property that each of the faces of I^{k+1} has $(k-1)$-dimensional edges, and when computing $\partial\circ\partial(I^{k+1})$, each of these edges appears as the edge of exactly two faces with opposite orientation! The proof given in Spivak is merely an analytical way of book-keeping this property.

Example 7.4.6 Example relating the integral of a two form to a classical surface integral.

Consider the two form

$$\omega(x,y,z) = P(x,y,z)\,dy\wedge dz + Q(x,y,z)\,dz\wedge dx + R(x,y,z)\,dx\wedge dy$$

in $\mathbb{R}^3$. Suppose $\vec\gamma : (u,v)\in I^2 \mapsto \mathbb{R}^3$ is a singular 2-cube. We defined $\int_{\vec\gamma}\omega$ as $\int_{I^2}\vec\gamma^*(\omega)$. Let's see a more concrete form of $\vec\gamma^*(\omega)$.

$$
\begin{aligned}
\vec\gamma^*(\omega) = P\circ\vec\gamma&\left(\frac{\partial y}{\partial u}du + \frac{\partial y}{\partial v}dv\right)\wedge\left(\frac{\partial z}{\partial u}du + \frac{\partial z}{\partial v}dv\right)\\
+Q\circ\vec\gamma&\left(\frac{\partial z}{\partial u}du + \frac{\partial z}{\partial v}dv\right)\wedge\left(\frac{\partial x}{\partial u}du + \frac{\partial x}{\partial v}dv\right)\\
+R\circ\vec\gamma&\left(\frac{\partial x}{\partial u}du + \frac{\partial x}{\partial v}dv\right)\wedge\left(\frac{\partial y}{\partial u}du + \frac{\partial y}{\partial v}dv\right)\\
=P\circ\vec\gamma&\left(\frac{\partial y}{\partial u}\frac{\partial z}{\partial v} - \frac{\partial y}{\partial v}\frac{\partial z}{\partial u}\right)du\wedge dv
\end{aligned}
$$

$$+ Q \circ \vec{\gamma} \left(\frac{\partial z}{\partial u} \frac{\partial x}{\partial v} - \frac{\partial z}{\partial v} \frac{\partial x}{\partial u} \right) du \wedge dv$$
$$+ R \circ \vec{\gamma} \left(\frac{\partial x}{\partial u} \frac{\partial y}{\partial v} - \frac{\partial x}{\partial v} \frac{\partial y}{\partial u} \right) du \wedge dv$$

Note that

$$D_u \vec{\gamma} \times D_v \vec{\gamma} = \left(\frac{\partial y}{\partial u} \frac{\partial z}{\partial v} - \frac{\partial y}{\partial v} \frac{\partial z}{\partial u}, \frac{\partial z}{\partial u} \frac{\partial x}{\partial v} - \frac{\partial z}{\partial v} \frac{\partial x}{\partial u}, \frac{\partial x}{\partial u} \frac{\partial y}{\partial v} - \frac{\partial x}{\partial v} \frac{\partial y}{\partial u} \right),$$

so

$$\vec{\gamma}^*(\omega) = (P \circ \vec{\gamma}, Q \circ \vec{\gamma}, R \circ \vec{\gamma}) \cdot (D_u \vec{\gamma} \times D_v \vec{\gamma})\, du \wedge dv,$$

and if we use ν to denote the unit vector in the direction of $D_u \vec{\gamma} \times D_v \vec{\gamma}$ (when it is not $\mathbf{0}$), then

$$\vec{\gamma}^*(\omega) = (P \circ \vec{\gamma}, Q \circ \vec{\gamma}, R \circ \vec{\gamma}) \cdot \nu \| D_u \vec{\gamma} \times D_v \vec{\gamma} \|\, du \wedge dv$$

and

$$\int_{I^2} \vec{\gamma}^*(\omega) = \int_{I^2} (P \circ \vec{\gamma}, Q \circ \vec{\gamma}, R \circ \vec{\gamma}) \cdot \nu \| D_u \vec{\gamma} \times D_v \vec{\gamma} \|\, du\, dv.$$

Recall that the integral of $\| D_u \vec{\gamma} \times D_v \vec{\gamma} \|$ gives the surface area of the parametric surface $\vec{\gamma} = \vec{\gamma}(u, v)$ over the parameter domain, and the integral on the right hand side above is identified as the surface integral

$$\int_{\vec{\gamma}(I^2)} (P, Q, R) \cdot \nu\, dA.$$

Because of this relation one often finds in older texts that $\int_S (P, Q, R) \cdot \nu\, dA$ is written as $\int_S P\, dydz + Q\, dzdx + R\, dxdy$ on a surface S.

If we look at the term involving P, Q or R individually, say, the term $P \circ \vec{\gamma} \left(\frac{\partial y}{\partial u} \frac{\partial z}{\partial v} - \frac{\partial y}{\partial v} \frac{\partial z}{\partial u} \right)$, then

$$\int_{I^2} P \circ \vec{\gamma} \left(\frac{\partial y}{\partial u} \frac{\partial z}{\partial v} - \frac{\partial y}{\partial v} \frac{\partial z}{\partial u} \right) dudv = \int_{\text{Projection into the } y,z \text{ plane of } \vec{\gamma}(I^2)} P\, dydz$$

according to the change of variables formula in integration, provided that $\frac{\partial y}{\partial u} \frac{\partial z}{\partial v} - \frac{\partial y}{\partial v} \frac{\partial z}{\partial u} > 0$. In other words, this integral can be treated as an integral in the y–z coordinate plane. When the sign is not specified, the integral takes into account of the orientation of the projection, such as in the case that $\vec{\gamma}$ is a parametrization for the round sphere that includes portions of both the upper and lower hemisphere, where $\frac{\partial y}{\partial u} \frac{\partial z}{\partial v} - \frac{\partial y}{\partial v} \frac{\partial z}{\partial u}$ will be positive in some region while negative in some other region, so the resulting integral reflects this. As a result, $\int_{\vec{\gamma}} \omega$ can be interpreted as projecting ω to each of the possible two dimensional coordinate planes and computing the integral of the projected component in the projected domain, taking into account of the orientation of the projection, and summing up these integrals.

For example, if $\vec{\gamma}$ is a parametrization for the upper half unit sphere centered at the origin, then its projection onto the y–z plane will be two-to-one, with the right half and left half carrying opposite signs in the Jacobian. Suppose the unit normal corresponding to the parametrization $\vec{\gamma}$ is outward with respect to the sphere. Since $\frac{\partial y}{\partial u} \frac{\partial z}{\partial v} - \frac{\partial y}{\partial v} \frac{\partial z}{\partial u}$ is the x-component of the normal, we know it is positive when $\vec{\gamma}$ lands on the right half of the sphere and is negative and when $\vec{\gamma}$ lands on the left half of the sphere. If $\omega = x\, dy \wedge dz$, then the x factor would also take opposite signs on the two halves of the sphere, so

$$\int_{\vec{\gamma}:x=\vec{\gamma}_1\geq 0} x\,dy\wedge dz = \int_{y^2+z^2\leq 1, z\geq 0} \sqrt{1-y^2-z^2}\,dydz,$$

$$\int_{\vec{\gamma}:x=\vec{\gamma}_1\leq 0} x\,dy\wedge dz = \int_{y^2+z^2\leq 1, z\geq 0} -\sqrt{1-y^2-z^2}(-1)\,dydz,$$

which results in

$$\int_{\vec{\gamma}} x\,dy\wedge dz = 2\int_{y^2+z^2\leq 1, z\geq 0} \sqrt{1-y^2-z^2}\,dydz,$$

provided the parametrization gives a positive Jacobian when $x>0$ (which is the case if $D_u\vec{\gamma}\times D_v$ points upward).

If one chooses to use $\vec{\gamma}(x,y)=(x,y,\sqrt{1-x^2-y^2})$ as a parametrization, then one could also use this parametrization to evaluate the integral directly, using

$$dy\wedge dz = dy\wedge(\frac{\partial z}{\partial x}dx+\frac{\partial z}{\partial y}dy) = -\frac{\partial z}{\partial x}dx\wedge dy,$$

and get

$$\int_{\vec{\gamma}} x\,dy\wedge dz = -\int_{x^2+y^2\leq 1} x\frac{\partial z}{\partial x}\,dx\,dy = \int_{x^2+y^2\leq 1}\frac{x^2}{\sqrt{1-x^2-y^2}}\,dx\,dy.$$

A third option to evaluate the above integral is to think of $\vec{\gamma}$ as part of ∂c for some singular 3-cube (or chain). One could imagine defining a 3-cube c mapping into the upper half unit ball such that its restriction to its top face gives rise to $\vec{\gamma}$, its restriction to its bottom face gives rise to a parametrization for the unit disk in the plane $z=0$, and its restriction to its lateral faces maps to the unit circle $x^2+y^2=1, z=0$, then

$$\int_{\partial c} x\,dy\wedge dz = \int_{\vec{\gamma}} x\,dy\wedge dz - \int_{\text{bottom}} x\,dy\wedge dz.$$

But on the bottom face, $z=0$, so $dy\wedge dz=0$ in the integral. On the other hand, we can apply the Stokes' Theorem to get

$$\int_{\vec{\gamma}} x\,dy\wedge dz = \int_c d(x\,dy\wedge dz) = \int_c dx\wedge dy\wedge dz.$$

A similar consideration for how to compute integrals of a 3-form leads us to conclude that $\int_c dx\wedge dy\wedge dz$ is the volume of the upper half unit sphere.

In higher dimensions, a surface may no longer have a notion of a normal vector, but the latter interpretation of projecting on each coordinate plane still applies.

Exercises

1. Define $c(u,v)=(u,v)$ for $(u,v)\in I^2$. For any differentiable 1-form $\omega = P\,dx+Q\,dy$ in I^2, work out $(\partial c)^*(\omega)$ and $d\,(c^*(\omega))$.

2. Let $c(u,v)=(f_1(u,v),f_2(u,v),f_3(u,v))$ be a differentiable map from I^2 to $\mathbb{R}^3$ and $\omega = P\,dx+Q\,dy+R\,dz$ be a differentiable 1-form in $\mathbb{R}^3$. Work out $(\partial c)^*(\omega)$ and $d\,(c^*(\omega))$.

3. Let m be an integer and $R > 0$. Define $c_{R,m}(t) = (R\cos(2\pi mt), R\sin(2\pi mt))$ as a map $I^1 \mapsto \mathbb{R}^2 \setminus \{(0,0)\}$. For any $r_1, r_2 > 0, r_1 \neq r_2$, construct a map $c : I^2 \mapsto \mathbb{R}^2 \setminus \{(0,0)\}$ such that $c_{r_2,m} - c_{r_1,m} = \partial c$ as a singular 2-chain.

4. If c is a singular 1-cube in $\mathbb{R}^2 - \{\mathbf{0}\}$ with $c(0) = c(1)$, show that there is an integer m such that $c - c_{1,m} = \partial c^2$ for some 2-chain c^2. Here $c_{R,m}(t) = R(\cos(2\pi mt), \sin(2\pi mt))$.

7.4.3 Closed and Exact Forms

Definition 7.4.7

A k-form ω is called closed if $d\omega = 0$; it is called exact if there exists a $(k-1)$-form η such that $\omega = d\eta$.

Note that, due to $d \circ d = 0$, any exact form must be closed.
Questions that we need to address include

- How do we check whether a form is closed or exact?
- Is every closed form also exact?

Since for any one form $\omega(\mathbf{x}) = \sum_{i=1}^m \omega_i(\mathbf{x})dx_i$ we have

$$d\omega(\mathbf{x}) = \sum_{j<i} \left(\frac{\partial \omega_i}{\partial x_j} - \frac{\partial \omega_j}{\partial x_i}\right) dx_j \wedge dx_i,$$

it's clear that $d\omega = 0$ iff $\frac{\partial \omega_i}{\partial x_j} = \frac{\partial \omega_j}{\partial x_i}$ for all i, j. These are $\frac{m(m-1)}{2}$ conditions on m components of ω.

For a two-form $\omega(\mathbf{x}) = \sum_{i<j} \omega_{ij} dx_i \wedge dx_j$,

$$d\omega(\mathbf{x}) = \sum_{i<j} d\omega_{ij} \wedge dx_i \wedge dx_j,$$

and if one examines the coefficient of $dx_i \wedge dx_j \wedge dx_k$ for some $i < j < k$, one gets

$$D_k\omega_{ij} + D_i\omega_{jk} - D_j\omega_{ik}.$$

So $d\omega = 0$ if and only if all the triple sums above are equal to zero. But these conditions are not that easy to comprehend. In dimension 4, a two form would have 6 components, and $d\omega = 0$ encodes 4 equations.

Based on our observation earlier that $\omega = d\eta$ for any one-form η would satisfy $d\omega = 0$, we get plenty of solutions this way, and the solutions are in terms of m arbitrary differentiable functions as components of η in regions of $\mathbb{R}^m$. Whether these provide all possible solutions is the second question raised above.The answer turns out to depend on the topology of the domain.

The one form $\omega = -\frac{y}{x^2+y^2}dx + \frac{x}{x^2+y^2}dy$ is closed on $\mathbb{R}^2 \setminus \{0\}$, but is not exact, for it were equal to df for some differentiable function f on $\mathbb{R}^2 \setminus \{0\}$, then for any 1-cube c in $\mathbb{R}^2 \setminus \{0\}$, $\int_c df = f(c(1)) - f(c(0))$ as the simplest form of Stokes' Theorem. As a consequence, $\int_c df = 0$ when $c(1) = c(0)$. But if we take $c(t) = (\cos(2\pi t), \sin(2\pi t))$, $0 \le t \le 1$, we find that $\int_c \omega = 2\pi$.

Definition 7.4.8

A set $A \subset \mathbb{R}^m$ is called star-shaped, if there exists some $O \in A$ such that for any $P \in A$ and any $t \in [0,1]$, $(1-t)O + tP \in A$.

Theorem 7.4.9 Poincarè Lemma.

Assume that A is an open star-shaped domain in $\mathbb{R}^m$, then any closed form on A is an exact form.

Proof. We may assume that A is star-shaped with respect to the origin. The key information we use is that $F(t,x) = tx$ for $(t,x) \in [0,1] \times A$ is a homotopy in A of $F(1,\cdot)$ and $F(0,\cdot)$. The heart of the argument is that there exists $I : \Lambda^{l+1}(TA) \mapsto \Lambda^l(TA)$ for $l = 1, \ldots, m$ such that

$$\omega = I(d\omega) + d\,(I(\omega)) \text{ for any } \omega \in \Lambda^k(TA). \tag{7.4.1}$$

As a result, when $d\omega = 0$, we find $\omega = d\,(I(\omega))$.

Let's assume that

$$\omega(\mathbf{x}) = \sum_{1 \le i_1 < \cdots < i_k \le m} \omega_{i_1 \cdots i_k}(\mathbf{x}) dx_{i_1} \wedge \cdots \wedge dx_{i_k}.$$

For each $t \in [0,1]$,

$$F(t,\cdot)^*(\omega) = \sum_{1 \le i_1 < \cdots < i_k \le m} t^k \omega_{i_1 \cdots i_k}(t\mathbf{x}) dx_{i_1} \wedge \cdots \wedge dx_{i_k}$$

is a k-form on A. Note that $F(1,\cdot)^*(\omega) = \omega$ and $F(0,\cdot)^*(\omega) = 0$.

Consideration for $D_t F(t,\cdot)^*(\omega)$ leads to the k-form on $[0,1] \times A$

$$F^*(\omega) = \sum_{1 \le i_1 < \cdots < i_k \le m} \omega_{i_1 \cdots i_k}(t\mathbf{x}) \left(t\,dx_{i_1} + x_{i_1} dt\right) \wedge \cdots \wedge \left(t\,dx_{i_k} + x_{i_k} dt\right).$$

$F^*(\omega)(\frac{\partial}{\partial t}, \cdot)$ defines a $(k-1)$-form on $[0,1] \times A$ by

$$F^*(\omega)(\frac{\partial}{\partial t}, \cdot) : (\mathbf{v}_1, \cdots, \mathbf{v}_{k-1}) \mapsto F^*(\omega)(\frac{\partial}{\partial t}, \mathbf{v}_1, \cdots, \mathbf{v}_{k-1}).$$

We define

$$I(\omega) = \int_0^1 F^*(\omega)(\frac{\partial}{\partial t}, \cdot)\, dt.$$

More concretely, $F^*(\omega)(\frac{\partial}{\partial t}, \cdot)$ is obtained from $F^*(\omega)$ by only keeping terms with one factor of $x_{i_l} dt$ and the remaining factors of $t dx_{i_j}$, then dropping the dt factor (as its action on $\frac{\partial}{\partial t}$ would be 1) and adjusting the sign according to the position of $x_{i_l} dt$, and integrating out the resulting expression in t to get

$$\begin{aligned} I(\omega) = \sum_{1 \le i_1 < \cdots < i_k \le m} \int_0^1 & t^{k-1} \omega_{i_1 \cdots i_k}(t\mathbf{x})\, dt\, \big(x_{i_1} dx_{i_2} \wedge \cdots \wedge dx_{i_k} \\ & - x_{i_2} dx_{i_1} \wedge dx_{i_3} \wedge \cdots \wedge dx_{i_k} + \cdots + (-1)^{k+1} x_{i_k} dx_{i_1} \wedge \cdots \wedge dx_{i_{k-1}}\big). \end{aligned}$$

It is then straightforward to verify that

$$I(d\omega) + d\,(I(\omega))$$

$$
\begin{aligned}
&= \sum_{1\le i_1<\cdots<i_k\le m} \int_0^1 \left(\sum_{l=1}^{m} t^k x_l D_l \omega_{i_1\cdots i_k}(t\mathbf{x}) + k t^{k-1}\omega_{i_1\cdots i_k}(t\mathbf{x}) \right) dt \\
&= \omega(\mathbf{x}).
\end{aligned}
$$

Below are the verifications.

$$
\begin{aligned}
dI(\omega) = & \sum_{1\le i_1<\cdots<i_k\le m} \Bigg\{ \sum_{l=1}^{m} \left(\int_0^1 t^k D_l\omega_{i_1\cdots i_k}(t\mathbf{x})\, dt \right) dx_l \wedge (x_{i_1} dx_{i_2} \wedge \cdots \wedge dx_{i_k} \\
& -x_{i_2} dx_{i_1} \wedge dx_{i_3} \wedge \cdots \wedge dx_{i_k} + \cdots + (-1)^{k+1} x_{i_k} dx_{i_1} \wedge \cdots \wedge dx_{i_{k-1}}) \\
& + \left(\int_0^1 t^{k-1}\omega_{i_1\cdots i_k}(t\mathbf{x})\, dt \right) (dx_{i_1} \wedge dx_{i_2} \wedge \cdots \wedge dx_{i_k} \\
& -dx_{i_2} \wedge dx_{i_1} \wedge dx_{i_3} \wedge \cdots \wedge dx_{i_k} + \cdots + (-1)^{k+1} dx_{i_k} \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_{k-1}}) \Big\} \\
= & \sum_{1\le i_1<\cdots<i_k\le m} \Bigg\{ \sum_{l=1}^{m} \left(\int_0^1 t^k D_l\omega_{i_1\cdots i_k}(t\mathbf{x})\, dt \right) dx_l \wedge (x_{i_1} dx_{i_2} \wedge \cdots \wedge dx_{i_k} \\
& -x_{i_2} dx_{i_1} \wedge dx_{i_3} \wedge \cdots \wedge dx_{i_k} + \cdots + (-1)^{k+1} x_{i_k} dx_{i_1} \wedge \cdots \wedge dx_{i_{k-1}}) \\
& + \left(\int_0^1 k t^{k-1}\omega_{i_1\cdots i_k}(t\mathbf{x})\, dt \right) dx_{i_1} \wedge dx_{i_2} \wedge \cdots \wedge dx_{i_k} \Bigg\},
\end{aligned}
$$

while

$$
\begin{aligned}
I(d\omega) = & \sum_{1\le i_1<\cdots<i_k\le m; 1\le l\le m} \Bigg\{ \left(\left(\int_0^1 t^k D_l\omega_{i_1\cdots i_k}(t\mathbf{x})\, dt \right) (x_l dx_{i_1} \wedge dx_{i_2} \wedge \cdots \wedge dx_{i_k} \right. \\
& + dx_l \left[-x_{i_1} dx_{i_2} \wedge \cdots \wedge dx_{i_k} + x_{i_2} dx_{i_1} \wedge dx_{i_3} \wedge \cdots \wedge dx_{i_k} \right. \\
& \left. + \cdots + (-1)^k x_{i_k} dx_{i_1} \wedge \cdots \wedge dx_{i_{k-1}} \right]) \Big\}
\end{aligned}
$$

Adding the above two terms completes our verification. ■

Remark 7.4.10

We used a specific homotopy $F(t,\cdot)$ in this proof for a star-shaped domain; this proof can be adapted to work for other homotopy. In particular, it can be used to show that if $F(t,\cdot)$ is a (differentiable) homotopy for $t \in [a,b]$, and ω is a closed form, then $F(b,\cdot)^*(\omega) - F(a,\cdot)^*(\omega)$ is exact, and often one can construct an $\eta = I(\omega)$ explicitly as described in the proof of Poincarè's Lemma so that $d\eta = F(b,\cdot)^*(\omega) - F(a,\cdot)^*(\omega)$.

Exercises

1. Let $\Theta(x,y) = \frac{-y\,dx + x\,dy}{x^2+y^2}$ be the 1-form on $\mathbb{R}^2 \setminus \{(0,0)\}$.

(a) Show that $\Theta(x,y)$ is closed on $\mathbb{R}^2 \setminus \{(0,0)\}$ but not exact in $\mathbb{R}^2 \setminus \{(0,0)\}$.

(b) Show also that $\Theta(x,y)$ is exact on $\mathbb{R}^2 \setminus \{(x,0) : x \le 0\}$.

(c) Suppose that ω is a 1-form on $\mathbb{R}^2 - \{\mathbf{0}\}$ such that $d\omega = 0$, prove that $\omega = \lambda\Theta + dg$ for some $\lambda \in \mathbb{R}$ and and $g : \mathbb{R}^2 - \{\mathbf{0}\} \mapsto \mathbb{R}$.

Hint. With $c_{R,1}(t) = (R\cos(2\pi t), R\sin(2\pi t))$ for $t \in I^1$, and $c^*_{R,1}(\omega)$ being closed on I^1, we can write

$$c^*_{R,1}(\omega) = \lambda_R dt + d(g_R),$$

for some function $g_R(t)$ satisfying $g_R(1) = g_R(0)$. Show that the number is independent of $R > 0$.

2. Consider the two-form

$$\omega = \frac{x\,dy \wedge dz + ydz \wedge dx + zdx \wedge dy}{(x^2 + y^2 + z^2)^{3/2}}$$

on $\mathbb{R}^3 \setminus \{\mathbf{0}\}$.

(a) Verify that $d\omega = 0$.

(b) Suppose that $c : (u, v) \in \mathbb{R}^2 \mapsto \mathbb{R}^3 \setminus \{\mathbf{0}\}$ is differentiable. Verify that

$$c^*(\omega) = \frac{c(u, v) \cdot (D_u c \times D_v c)}{|c(u, v)|^3} du \wedge dv.$$

(c) Suppose that $c\Big|_{I^2}$ is injective and $|D_u c \times D_v c| > 0$ so that $c(I^2)$ is a differentiable surface in $\mathbb{R}^3 \setminus \{\mathbf{0}\}$. Show that

$$\int_{c(I^2)} \frac{(x, y, z) \cdot \mathbf{n}(x, y, z)}{(x^2 + y^2 + z^2)^{3/2}}\, dA = \int_{I^2} c^*(\omega),$$

where

$$\mathbf{n}(c(u, v)) = \frac{D_u c \times D_v c}{|D_u c \times D_v c|}$$

is a unit normal to the parametric surface $c(u, v)$ at $c(u, v)$.

(d) Choose the spherical polar coordinate parametrization s for the sphere $x^2+y^2+z^2 = R^2$ defined for $(\phi, \theta) \in [0, \pi]\times[0, 2\pi]$ to show that $D_u s\times D_v s$ is an outward normal and that $\int_s \omega = 4\pi$.

(e) Show that for any continuous 1-form η in $\mathbb{R}^3 \setminus \{\mathbf{0}\}$, $\int_{\partial s} \eta = 0$.

(f) Show that there does not exist a 1-form η in $\mathbb{R}^3 \setminus \{\mathbf{0}\}$ such that $d\eta = \omega$ in $\mathbb{R}^3 \setminus \{\mathbf{0}\}$.

References

[1] P. Lax: *Rethinking the Lebesgue Integral*, The American Mathematical Monthly, Dec., 2009, Vol. 116, No. 10 (Dec., 2009), pp. 863-881.

[2] James Munkres: *Analysis on Manifolds*, 1991, Addison-Wesley Publishing Company.

[3] Walter Rudin: *Principles of Mathematical Analysis*, 3rd edition (1976), McGraw-Hill.

[4] Michael Spivak: *Calculus on Manifolds*, 1965, Addison-Wesley Publishing Company.

[5] Min-qiang Zhou: *Theory of Functions of A Real Variable* (in Chinese), 3rd edition (2001), Peking University Press.

[6] Vladimir A. Zorich: *Mathematical Analysis I, II*[1], 2nd edition (2016), Springer-Verlag Berlin Heidelberg.

[1] `libguides.rutgers.edu/math/ebooks`