Advanced Calculus

Feb - June 2024 Prof. Dr. Marc Paolella

> I advise my students to listen carefully the moment they decide to take no more mathematics courses. They might be able to hear the sound of closing doors.
> (James Caballero)

> One difficult decision relates to how much of an effort one should make to acquire basic technique, for example in mathematics and probability theory. One does not wish to be always training to run the race but never running it; however, we do need to train. (E. J. Hannan, 1992)

The beginning masters students at UZH majoring in Business, Data Science, Economics, or Finance, have had a basic course in univariate calculus, and this, during their bachelor studies. Having taught my beginning master's level class in probability theory for 20 years in a row, I know very well the average student level of understanding in calculus, linear algebra, and basic mathematics. It is not very high. If the student's goal is to get more involved in advanced (micro- or macro-) economics, econometrics, quantitative risk management, asset pricing, probability theory, higher level computational-based statistical methods and machine learning, hardcore mathematical finance and financial engineering, etc., then he or she will need to have a much stronger level of mathematics than the typical rudimentary level. Filling this gap is the purpose of this course. We will use this document, put together precisely for this course, in conjunction (time-permitting) with other materials (see below). This document began from the large math appendix of my book Fundamental Probability: A Computational Approach (2006), along with some material from its chapters 1 and 9; and got (corrected and) enormously expanded.

The term "advanced calculus" is often used synonymously with a course in real analysis that focuses on the multivariate case and chronologically follows, obviously, a first course in real analysis. In our course, we will in fact also cover the univariate case, reviewing some of the material covered in the first course on analysis, but with emphasis on "computable, tangible things", as opposed to the more abstract notions in real analysis required to understand more advanced analysis and measure theory, notably compactness (see the quote at the beginning of §4.1 regarding its relevance). As such, we will not refer to metric spaces, and not invoke the concepts of sequential and topological compactness, subsequences, and the Bolzano-Weierstrass and Heine-Borel theorems. These are covered in my univariate course, Real Analysis I.

The next goal is to cover the important concepts of series of numbers, and, even more relevantly, series of functions, reaching the immensely important topic of Taylor series, which we do in both the univariate and multivariate case. One of the goals of the course is to offer much practice by way of a large number of worked examples, such as "trickier" univariate Riemann integrals. Then we investigate a select set of topics associated with multivariate calculus, notably vectors and linear algebra, (partial and total) differentiation, and multivariate Riemann integration (importantly, Fubini's theorem, exchange of derivative and integral, and Leibnitz' rule).

Depending on time and interest, we could also cover the (univariate) Riemann-Stieltjes integral, using, say, Stoll (2021, §6.5). Other possible topics include Lagrange multipliers, and Fourier analysis and transforms. For the latter, I have prepared a separate document, being a compilation of chapters from three other books, namely Paolella, Stoll, and Kuttler.

Differing from a typical calculus or advanced calculus course, we will start in the first chapter with material that is highly relevant in general, and notably so with probability theory, namely some more sophisticated combinatorics, generalized binomial theorems, gamma and beta functions, Wallis' product and Stirling's approximation, and numerous non-trivial examples invoking this material. Appendix 6.2 contains material on the so-called digamma and polygamma functions. A further nod to probability theory is showing several ways of computing the univariate integral associated with the Gaussian distribution (Examples 5.20 and 5.21); and §5.7, covering some more elaborate examples of multivariate transformations of random variables.

Throughout the document, many, but not all, results are proven. Understanding proofs is essential in this course, but the main emphasis is on practical examples of a nontrivial nature, going well beyond the trivial examples in a first, undergraduate, course in calculus for students in the social sciences. I occasionally refer to results coming later in the document, e.g., Example 1.8 involves a Riemann integral, and I refer to its linearity property, as stated later, in §2.4.1. The reader is not expected to jump ahead and understand that material; it is there for reference. In my experience, having perused and studied numerous excellent mathematics books, this approach is quite common, because it is nearly unavoidable, notably in presentations such as this one, which cover a variety of topics. The key is to not do it too often!

It is worth mentioning that, in addition to the material herein, students aspiring to learn one or more of the aforementioned topics, e.g., mathematical finance, quant risk management, etc., will also require the *sine qua non* (indispensable and essential action, condition, or ingredient) of linear algebra, advanced statistical methods, measure theory and the Lebesgue integral, and (measure-theoretic) probability theory; not to mention computer coding skills. This set is nicely captured in the preface of the well-received (2004, corrected 2009) book Convex Optimization, by Boyd and Vandenberghe: "The only background required of the reader is a good knowledge of advanced calculus and linear algebra. If the reader has seen basic mathematical analysis (e.g., norms, convergence, elementary topology), and basic probability theory, he or she should be able to follow every argument and discussion in the book."

Lecture syllabus: Date and page numbers: As of March 12, add 10 to all the pages, due to me having added material before page 60.

Feb 20 1 - 10	Apr 16 141 - 150
Feb 22 11 - 20	Apr 18 151 - 160
Feb 27 21 - 30	Apr 23 161 - 170
Feb 29 31 - 40	Apr 25 171 - 180
Mar 5 41 - 50	Apr 30 181 - 190
Mar 7 51 - 60	May 2 191 - 200
Mar 12 61 - 70	May 7 201 - 210
Mar 14 71 - 80	May 14 211 - 220
Mar 19 81 - 90	May 16 221 - 230
Mar 21 91 - 100	May 21 231 - 240
Mar 26 101 - 110	May 23 241 - 250
Mar 28 111 - 120	May 28 251 - 260
Apr 9 121 - 130	May 30 261 - 294
Apr 11 131 - 140	

Contents

1 Preliminaries **1**
- 1.1 Sets, Functions, and Fundamental Inequalities . . . 1
- 1.2 Binomial and Generalized Binomial Theorems . . . 6
- 1.3 Gamma and Beta Functions . . . 18

2 Univariate Calculus **25**
- 2.1 Limits and Continuity . . . 25
- 2.2 Differentiation . . . 33
  - 2.2.1 Definitions and Techniques . . . 33
  - 2.2.2 Trigonometric Functions . . . 38
  - 2.2.3 Mean Value Theorem and Function Extreme Points . . . 42
  - 2.2.4 Exponential and Logarithm . . . 49
- 2.3 Convexity . . . 55
- 2.4 Integration . . . 63
  - 2.4.1 Definitions, Existence, and Properties . . . 63
  - 2.4.2 Fundamental Theorem of Calculus . . . 68
  - 2.4.3 Improper Integrals . . . 82
- 2.5 Series . . . 95
  - 2.5.1 Definitions and Examples . . . 95
  - 2.5.2 Tests for Convergence and Divergence . . . 103
  - 2.5.3 Wallis' Product and Stirling's Approximation . . . 113
  - 2.5.4 Cauchy Product . . . 117
  - 2.5.5 Sequences and Series of Functions . . . 119
  - 2.5.6 Power and Taylor Series . . . 134

3 Some Relevant Linear Algebra **141**
- 3.1 (Hyper-)planes, Vector-Parametric and Cartesian Equations . . . 141
- 3.2 Projection . . . 150
  - 3.2.1 Shifrin and Adams, Linear Algebra: A Geometric Approach . . . 150
  - 3.2.2 Flanigan and Kazdan, Calculus Two: Linear and Non-linear Functions 153
  - 3.2.3 Lang, Calculus of Several Variables . . . 157
- 3.3 Determinants and the Cross Product . . . 163
- 3.4 More Advanced Linear Algebra: Projection and Least Squares . . . 167
  - 3.4.1 Inner Product Spaces and Gram Matrices . . . 167
  - 3.4.2 Orthogonal and Orthonormal Bases . . . 171
  - 3.4.3 Gram–Schmidt, Orthogonal Matrices, QR Factorization . . . 173
  - 3.4.4 Orthogonal Projections and Orthogonal Subspaces . . . 180
  - 3.4.5 Least Squares Minimization . . . 184

4 Multivariate Calculus: Differentiation, Tangent Maps, and Taylor Series 191
4.1 Sequences, Limits, Functions, and Continuity . . . 191
4.2 Partial Derivatives and the Gradient . . . 202
4.3 Differentiability and Tangent Maps . . . 205
4.4 Higher Order Partial Derivatives . . . 219
4.5 Directional Derivatives and the Multivariate MVT . . . 223
4.6 The Jacobian and the Chain Rule . . . 230
4.6.1 The Jacobian . . . 230
4.6.2 The Chain Rule . . . 233
4.6.3 The Mean Value Theorem (MVT) . . . 238
4.7 Higher Order Derivatives and Taylor Series . . . 239
4.8 Local Approximation of Real-Valued Multivariate Functions . . . 241
4.9 Approximating Nonlinear Mappings By Linear Mappings . . . 252
4.9.1 Derivative Matrix and Differential . . . 252
4.9.2 The Chain Rule . . . 255
4.9.3 Directional Derivatives . . . 259

5 Multivariate Integration 260
5.1 Definitions, Existence, and Properties . . . 260
5.2 Bounded Sets, Jordan Measure, Volume Zero, and Lebesgue Measure Zero . . . 265
5.2.1 Introduction and Useful Results . . . 265
5.2.2 Bounded Sets, Jordan Measure, Volume Zero . . . 266
5.2.3 Lebesgue Measure Zero . . . 269
5.3 Exchange of Derivative and Integral . . . 274
5.4 Fubini's Theorem . . . 276
5.5 Leibniz' Rule . . . 281
5.6 Integral Transformations, Polar and Spherical Coordinates . . . 284
5.7 Multivariate Transformations for Random Variables . . . 288

6 Appendices 295
6.1 Further Material on the Gamma Function . . . 295
6.2 The Digamma and Polygamma Functions . . . 300
6.3 Banach's Matchbox Problem . . . 303

1 Preliminaries

> The point of view that "natural number" cannot be defined would be contested by many mathematicians who would maintain that the concept of "set" is more primitive than that of "number" and who would use it to define "number". Others would contend that the idea of "set" is not at all intuitive and would contend that, in particular, the idea of an *infinite* set is very nebulous. They would consider a definition of "number" in terms of sets to be an absurdity because it uses a difficult and perhaps meaningless concept to define a simple one.
>
> (Harold M. Edwards, 1994, p. 461)

1.1 Sets, Functions, and Fundamental Inequalities

It turns out that, mathematically speaking, a precise definition of *set* is problematic. For our purposes, it can be thought of simply as a well-defined collection of objects. This intuitive description cannot be a definition, because the word "collection" is nothing but a synonym for the word set. Nevertheless, in all contexts considered herein, the notion of set will be clear. For example, if $A = \{n \in \mathbb{N} : n < 7\}$, then A is the set of positive integers less than 7, or $A = \{1, 2, \ldots, 6\}$. If a is contained in A, then we write $a \in A$; otherwise, $a \notin A$. A set without any objects is called the *empty set* and is denoted $\emptyset$. A set with exactly one element is a *singleton set*.

Let A and B be two sets. The following handful of basic set operations will be used repeatedly throughout:

- the *intersection* of two sets, "A and B" (or "A intersect B"), denoted $A \cap B$. Each element of $A \cap B$ is contained in A, and contained in B; $A \cap B = \{x : x \in A, x \in B\}$.

- the *union* of two sets, "A or B" (or "A union B"), denoted $A \cup B$. An element of $A \cup B$ is either in A, or in B, or in both.

- set *subsets*, "A is a subset of B" or "A is contained in B" or "B contains A", denoted $A \subset B$ or $B \supset A$. If every element contained in A is also in B, then $A \subset B$. Like the ordering symbols $\leq$ and $<$ for the real numbers, it is sometimes useful (if not more correct) to use the notation $A \subseteq B$ to indicate that A and B could be equal, and reserve $A \subset B$ to indicate that A is a *proper subset* of B, i.e., $A \subseteq B$ but $A \neq B$; in words, that there is at least one element in B that is not in A. Only when this distinction is important will we use $\subseteq$. Also, $\emptyset$ is a subset of every set.

- set *equality*, "$A = B$", which is true if and only if $A \subset B$ and $B \subset A$. To prove two sets are equal, we prove $A \subset B$, and $B \subset A$.

- the *difference*, or *relative complement*, "B setminus A", denoted $B \setminus A$ or, sometimes authors write $B - A$. It is the set of elements contained in B but not in A.

- If the set B is clear from the context, then it need not be explicitly stated, and the set difference $B \setminus A$ is written as A^c, which is the *complement* of A. Thus, we can write $B \setminus A = B \cap A^c$.

- the *product* of two sets, A and B, consists of all ordered pairs (a, b), such that $a \in A$ and $b \in B$; it is denoted $A \times B$.

The first four previous set operations are extended to more than two sets in a natural way, i.e., for intersection, if $a \in A_1 \cap A_2 \cap \cdots \cap A_n$, then a is contained in each of the A_i, and is abbreviated by $a \in \bigcap_{i=1}^n A_i$. A similar notation is used for union. To illustrate this for subsets, let $A_n = [1/n, 1]$, $n \in \{1, 2, \ldots\}$, i.e., $A_1 = \{1\}$ and $A_2 = [1/2, 1] = \{x : 1/2 \le x \le 1\}$. Then $A_1 \subset A_2 \subset \cdots$, and $\bigcup_{n=1}^\infty = (0, 1] = \{x : 0 < x \le 1\}$. In this case, the A_n are said to be *monotone increasing*. If sets A_i are monotone increasing, then

$$\lim_{i\to\infty} A_i = \bigcup_{i=1}^\infty A_i. \tag{1.1}$$

Similarly, the sets A_i are *monotone decreasing* if $A_1 \supset A_2 \supset \cdots$, in which case

$$\lim_{i\to\infty} A_i = \bigcap_{i=1}^\infty A_i. \tag{1.2}$$

We will also need basic familiarity with the following sets: $\mathbb{N} = \{1, 2, \ldots\}$ is the set of all natural numbers; $\mathbb{Z} = \{0, 1, -1, 2, -2, \ldots\}$ is the set of all integers or Zahlen (German for number); $\mathbb{Q} = \{m/n,\, m \in \mathbb{Z},\, n \in \mathbb{N}\}$ is the set of all rational numbers (quotients); $\mathbb{R}$ is the set of all real numbers; $\mathbb{C}$ is the set of complex numbers, and $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}$. For convenience and clarity, we also define $\mathbb{R}_{>0} = \{x : x \in \mathbb{R},\, x > 0\}$, $\mathbb{R}_{\ge 1} = \{x : x \in \mathbb{R},\, x \ge 1\}$, etc.; if only a range is specified, then the real numbers are assumed, e.g., $x > 0$ is the same as $x \in \mathbb{R}_{>0}$. Also, we take $\mathbb{X} := \mathbb{R} \cup \{-\infty, \infty\}$, which is the *extended real line*. Letting $a \in \mathbb{R}$, properties of $\mathbb{X}$ include $\infty + \infty = \infty + a = \infty$, $a \cdot \infty = \operatorname{sgn}(a) \cdot \infty$, but $\infty - \infty$, ∞/∞, etc., are undefined, as remains $0/0$.

We make use of the common abbreviations $\exists$ ("there exists"), $\nexists$ ("there does not exist"), $\Rightarrow$ ("implies"), iff (if and only if) and $\forall$ ("for all" or, better, "for each"; see Pugh, 2002, p. 5). As an example, $\forall x \in (0, 1)$, $\exists\, y \in (x, 1)$. Also, the notation "$A := B$" means that A, or the lhs (left hand side) of the equation, is defined to be B, or the rhs (right hand side).

Sets obey certain rules, such as $A \cup A = A$ (idempotent); $(A \cup B) \cup C = A \cup (B \cup C)$ and $(A \cap B) \cap C = A \cap (B \cap C)$ (associative); $A \cup B = B \cup A$ and $A \cap B = B \cap A$ (commutative);

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

(distributive); $A \cup \emptyset = A$ and $A \cap \emptyset = \emptyset$ (identity); and $(A^c)^c = A$ (involution). Less obvious are De Morgan's laws, after Augustus De Morgan (1806–1871), which state that $(A \cup B)^c = A^c \cap B^c$ and $(A \cap B)^c = A^c \cup B^c$. More generally,

$$\left(\bigcup_{n=1}^\infty A_n \right)^c = \bigcap_{n=1}^\infty A_n^c \quad \text{and} \quad \left(\bigcap_{n=1}^\infty A_n \right)^c = \bigcup_{n=1}^\infty A_n^c. \tag{1.3}$$

Example 1.1 *Let* $B_i := A_i \setminus [A_i \cap (A_1 \cup A_2 \cup \cdots \cup A_{i-1})]$. *We wish to demonstrate that*

$$B_i = A_i \setminus (A_1 \cup \cdots \cup A_{i-1}), \quad i \ge 2.$$

It is useful to take $i = 2$, *and draw a Venn diagram, confirming the result in this first case. The general proof is an excuse to practice using basic set theory relations.*

Use the above rules for sets to get

$$\begin{aligned}
B_i &= A_i \setminus [A_i \cap (A_1 \cup A_2 \cup \cdots \cup A_{i-1})] = A_i \cap [A_i \cap (A_1 \cup A_2 \cup \cdots \cup A_{i-1})]^c \\
&= A_i \cap [(A_i \cap A_1) \cup (A_i \cap A_2) \cup \cdots \cup (A_i \cap A_{i-1})]^c \\
&= A_i \cap [(A_i \cap A_1)^c \cap (A_i \cap A_2)^c \cap \cdots \cap (A_i \cap A_{i-1})^c] \\
&= A_i \cap (A_i^c \cup A_1^c) \cap (A_i^c \cup A_2^c) \cap \cdots \cap \left(A_i^c \cup A_{i-1}^c\right) \\
&= [A_i \cap (A_i^c \cup A_1^c)] \cap [A_i \cap (A_i^c \cup A_2^c)] \cap \cdots \cap \left[A_i \cap \left(A_i^c \cup A_{i-1}^c\right)\right] \\
&= [(A_i \cap A_i^c) \cup (A_i \cap A_1^c)] \cap \cdots \cap \left[(A_i \cap A_i^c) \cup \left(A_i \cap A_{i-1}^c\right)\right] \\
&= (A_i \cap A_1^c) \cap \cdots \cap \left(A_i \cap A_{i-1}^c\right) = A_i \cap \left(A_1^c \cap \cdots \cap A_{i-1}^c\right) \\
&= A_i \cap (A_1 \cup \cdots \cup A_{i-1})^c = A_i \setminus (A_1 \cup \cdots \cup A_{i-1}) .
\end{aligned}$$

■

Two sets are *disjoint*, or *mutually exclusive*, if $A \cap B = \emptyset$, i.e., they have no elements in common. A set J is an *indexing set* if it contains a set of indices, usually a subset of $\mathbb{N}$, and is used to work with a group of sets A_i, where $i \in J$. If A_i, $i \in J$, are such that $\bigcup_{i \in J} A_i \supset \Omega$, then they are said to *exhaust*, or (form a) *cover* (for) the set Ω. If sets A_i, $i \in J$, are nonempty, mutually exclusive and exhaust Ω, then they (form a) *partition* (of) Ω.

Example 1.2 *Let A_i be monotone increasing sets, i.e., $A_1 \subset A_2 \subset \cdots$. Define $B_1 := A_1$ and $B_i := A_i \setminus (A_1 \cup A_2 \cup \cdots \cup A_{i-1})$. We wish to show that, for $n \in \mathbb{N}$,*

$$\bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} B_i. \tag{1.4}$$

The B_i are clearly disjoint from their definition and such that B_i is the "marginal contribution" of A_i over and above that of $(A_1 \cup A_2 \cup \cdots \cup A_{i-1})$, which follows because the A_i are monotone increasing. Thus, $B_i = A_i \setminus A_{i-1} = A_i \cap A_{i-1}^c$. If $\omega \in \bigcup_{i=1}^n A_i$, then, because the A_i are increasing, either $\omega \in A_1$ (and, thus, in all the A_i) or there exists a value $j \in \{2, \ldots, n\}$ such that $\omega \in A_j$ but $\omega \notin A_i$, $i < j$. It follows from the definition of the B_i that $\omega \in B_j$ and thus in $\bigcup_{i=1}^n B_i$, so that (i) $\bigcup_{i=1}^n A_i \subset \bigcup_{i=1}^n B_i$.

Likewise, if $\omega \in \bigcup_{i=1}^n B_i$, then, as the B_i are disjoint, ω is in exactly one of the B_i, say B_j, $j \in \{1, 2, \ldots, n\}$. From the definition of B_j, $\omega \in A_j$, so $\omega \in \bigcup_{i=1}^n A_i$, so that (ii) $\bigcup_{i=1}^n B_i \subset \bigcup_{i=1}^n A_i$. Together, (i) and (ii) imply that $\bigcup_{i=1}^n A_i = \bigcup_{i=1}^n B_i$.

Also, for $i > 1$,

$$\begin{aligned}
B_i &= A_i \setminus (A_1 \cup A_2 \cup \cdots \cup A_{i-1}) = A_i \cap (A_1 \cup A_2 \cup \cdots \cup A_{i-1})^c \\
&= A_i A_1^c A_2^c \cdots A_{i-1}^c = A_i A_{i-1}^c,
\end{aligned}$$

where the last equality follows from $A_j = \bigcup_{n=1}^j A_n$ (because the A_i are monotone increasing) and, thus, $A_j^c = \bigcap_{n=1}^j A_n^c$. ■

For $a, b \in \mathbb{R}$ with $a \leq b$, the interval $(a, b) = \{x \in \mathbb{R} : a < x < b\}$ is said to be an *open interval*, while $[a, b] = \{x \in \mathbb{R} : a \leq x \leq b\}$ is a *closed interval*. In both cases, the interval has *length* $b - a$. For a set $S \subset \mathbb{R}$, the set of open intervals $\{O_i\}$, for $i \in J$ with J an indexing set, is an *open cover* of S if $\bigcup_{i \in J} O_i$ covers S, i.e., if $S \subset \bigcup_{i \in J} O_i$. Let $S \subset \mathbb{R}$ be such that there exists an open cover $\bigcup_{i \in N} O_i$ of S with a finite or countably infinite number of intervals. Denote the length of each O_i as $\ell(O_i)$. If $\forall \epsilon > 0$, there exists a cover $\bigcup_{i \in N} O_i$ of S such that

$$\sum_{i=1}^{\infty} \ell(O_i) < \epsilon, \tag{1.5}$$

then S is said to have *measure zero*. See also §5.2.3.

For our purposes, the most important set with measure zero is any set with a finite or countable number of points. For example, if f and g are functions with domain $I = (a, b) \in \mathbb{R}$, where $a < b$, and such that $f(x) = g(x)$ for all $x \in I$ except for a finite or countably infinite number of points in I, then we say that f **and** g **differ on** I **by a set of measure zero**. As an example from probability, if U is a continuous uniform random variable on $[0, 1]$, then the event that $U = 1/2$ is not impossible, but it has probability zero, because the point $1/2$ has measure zero, as does any finite collection of points, or any countably infinite set of points on $[0, 1]$, e.g., $\{1/n,\ n \in \mathbb{N}\}$.

Let S be a nonempty subset of $\mathbb{R}$. We say S has an *upper bound* M if $x \leq M\ \forall x \in S$, in which case S is *bounded above* by M. Note that, if S is bounded above, then it has infinitely many upper bounds. A fundamental property of $\mathbb{R}$ not shared by $\mathbb{Q}$ is that, if S is a nonempty set that has an upper bound M, then S possesses a unique *least upper bound*, or *supremum*, denoted $\sup S$. That is, $\exists U \in \mathbb{R}$ such that U is an upper bound of S, and such that, if V is also an upper bound of S, then $V \geq U$. If S is not bounded above, then $\sup S = \infty$. Also, $\sup \emptyset = -\infty$. Similar terminology applies to the *greatest lower bound*, or *infimum* of S, denoted $\inf S$. For example, let $S = \{1/n : n \in \mathbb{N}\}$. Then $\max S = \sup S = 1$ and $\inf S = 0$, but S has no minimum value. Next, let S consist of the truncated values of $\sqrt{2}$ with $n \in \mathbb{N}$ decimal places, i.e., $S = \{1.4, 1.41, 1.414, 1.4142, \ldots\}$. Then $S \subset \mathbb{Q}$ but $\sup S = \sqrt{2} \notin \mathbb{Q}$.[1]

A *relation* between A and B is a subset of $A \times B$. If a relation f is such that, for each $a \in A$, there is one and only one $b \in B$ such that $(a, b) \in f$, then f is also a *mapping*; one writes $f : A \to B$ and $b = f(a)$, with A referred to as the *domain* and B as the *codomain* or *target*. When f is plotted on the plane in the standard fashion, i.e., with A on the horizontal axis and B on the vertical axis, then a mapping satisfies the "vertical line test". The subset of the codomain given by $\{b \in B : \exists a \in A \text{ with } f(a) = b\}$ is the *range* or *image* of f. For some subset $C \subset B$, the *pre-image* of C is the subset of the domain given by $\{a \in A : f(a) \in C\}$.

A mapping with codomain $B = \mathbb{R}$ is a real-valued *function*. Let f be a function with domain A and let $I \in A$ be an interval. If f is such that, $\forall a, b \in A$, $a < b \Rightarrow f(a) < f(b)$, then f is *strictly increasing* on I. Likewise, if $a < b \Rightarrow f(a) \leq f(b)$, then f is *(weakly) increasing*. The terms *strictly decreasing* and *(weakly) decreasing* are similarly defined. A function that is either increasing or decreasing is said to be *monotone*, while a function that is either strictly increasing or strictly decreasing is *strictly monotone*.

The mapping $f : A \to B$ is *injective* or *one–to–one* if $f(a_1) = f(a_2)$ implies $a_1 = a_2$ (that is, if a plot of f satisfies the "horizontal line test"). A mapping is *surjective* or *onto* if the range is the (whole) codomain, and *bijective* if it is injective and surjective. If $f : A \to B$ is bijective, then the *inverse mapping* $f^{-1} : B \to A$ is bijective such that $f^{-1}(b)$ is the (unique) element in A such that $f(a) = b$. For mappings $f : A \to B$ and $g : B \to C$, the *composite mapping*, denoted $g \circ f : A \to C$, maps an element $a \in A$ to $g(f(a))$. Some thought confirms that, if f and g are injective, then so is $g \circ f$, and if f and g are surjective, then so is $g \circ f$; thus, if f and g are bijective, then so is $g \circ f$.

If $a \in \mathbb{R}$, then the *absolute value* of a is denoted by $|a|$, and is equal to a if $a \geq 0$, and $-a$ if $a < 0$. Clearly, $a \leq |a|$ and, $\forall a, b \in \mathbb{R}$, $|ab| = |a|\,|b|$. Observe that, for $b \in \mathbb{R}_{>0}$, the inequality $-b < a < b$ is equivalent to $|a| < b$, and, similarly,

$$-b \leq a \leq b \quad \Leftrightarrow \quad |a| \leq b. \tag{1.6}$$

[1] Observe that any element in $\mathbb{R}$ can be arbitrarily closely approximated by an element in $\mathbb{Q}$, which is the informal description of saying that $\mathbb{Q}$ is *dense* in $\mathbb{R}$. This is of enormous importance when actually working with numerical values in an (unavoidably) finite precision computing world.

Theorem (Triangle Inequality): $\forall x, y \in \mathbb{R}$,

$$|x+y| \leq |x| + |y| \,. \tag{1.7}$$

Proof: Square both sides to get

$$|x+y|^2 = (x+y)^2 = x^2 + 2xy + y^2 \quad \text{and} \quad (|x| + |y|)^2 = x^2 + 2\,|x|\,|y| + y^2.$$

Note $xy \leq |xy| = |x|\,|y|$. Alternatively, note that, $\forall a \in \mathbb{R}$, $-|a| \leq a \leq |a|$, so adding $-|x| \leq x \leq |x|$ to $-|y| \leq y \leq |y|$ gives $-(|x| + |y|) \leq x + y \leq |x| + |y|$, which, from (1.6) with $a = x + y$ and $b = |x| + |y|$, is equivalent to $|x + y| \leq |x| + |y|$. Also, with $z = -y$, the triangle inequality states that, $\forall x, z \in \mathbb{R}$, $|x - z| \leq |x| + |z|$.

Theorem (Reverse Triangle Inequality): $\forall a, b \in \mathbb{R}$,

$$||a| - |b|| \leq |a + b|, \quad \text{or} \quad ||a| - |b|| \leq |b - a|. \tag{1.8}$$

Proof: Write $|a| = |(a + b) + (-b)| \leq |a + b| + |b|$, or $|a| - |b| \leq |a + b|$. Switching a and b gives $|b| - |a| \leq |a + b|$ or $||a| - |b|| \leq |a + b|$. Replacing b with $-b$ gives $||a| - |b|| \leq |a - b| = |b - a|$.

Theorem (Cauchy-Schwarz Inequality): For any points $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ in $\mathbb{R}^n$, $n \in \mathbb{N}$,

$$|x_1 y_1 + \cdots + x_n y_n| \leq (x_1^2 + \cdots + x_n^2)^{1/2} (y_1^2 + \cdots + y_n^2)^{1/2}. \tag{1.9}$$

It is named after Augustin Louis Cauchy (1789–1857) and Hermann Schwarz (1843–1921), and also referred to as Cauchy's inequality or the Schwarz inequality. It was first published by Cauchy in 1821.

Proof: Let $f(r) = \sum_{i=1}^n (r x_i + y_i)^2 = Ar^2 + Br + C$, where $A = \sum_{i=1}^n x_i^2$, $B = 2\sum_{i=1}^n x_i y_i$ and $C = \sum_{i=1}^n y_i^2$. As $f(r) \geq 0$, the quadratic $Ar^2 + Br + C$ has one or no real roots, so that its discriminant $B^2 - 4AC \leq 0$, i.e., $B^2 \leq 4AC$ or, substituting, $\left(\sum_{i=1}^n x_i y_i\right)^2 \leq \left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n y_i^2\right)$, which is (1.9) after taking square roots.

The Cauchy-Schwarz inequality is used to show the generalization of (1.7):

Theorem (Triangle Inequality): For any points $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ in $\mathbb{R}^n$, $n \in \mathbb{N}$,

$$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| \,, \tag{1.10}$$

where

$$\|\mathbf{x}\| = \sqrt{x_1^2 + \cdots + x_n^2} \tag{1.11}$$

is the *norm* of $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$.

Proof: Using the above notation for A, B and C,

$$\|\mathbf{x} + \mathbf{y}\|^2 = \sum_{i=1}^n (x_i + y_i)^2 = \sum_{i=1}^n x_i^2 + 2\sum_{i=1}^n x_i y_i + \sum_{i=1}^n y_i^2 = A + B + C$$

and, as $B^2 \leq 4AC$, $A + B + C \leq A + 2\sqrt{AC} + C = \left(\sqrt{A} + \sqrt{C}\right)^2$. Taking square roots gives $\|\mathbf{x} + \mathbf{y}\| = \sqrt{A + B + C} \leq \sqrt{A} + \sqrt{C} = \|\mathbf{x}\| + \|\mathbf{y}\|$.

Remark: While we will not make use of this, it is worth mentioning that the Cauchy-Schwarz inequality can be generalized to *Hölder's inequality*: Let $p, q \in \mathbb{R}_{>1}$ be such that $(p-1)(q-1) = 1$ (or, equivalently $p^{-1} + q^{-1} = 1$). Then

$$|x_1 y_1 + \cdots + x_n y_n| \leq \left(|x_1|^p + \cdots + |x_n|^p\right)^{1/p} \left(|y_1|^q + \cdots + |y_n|^q\right)^{1/q},$$

while the triangle inequality can be generalized to *Minkowski's inequality*:

$$\left(\sum_{i=1}^{n} |x_i + y_i|^p\right)^{1/p} \leq \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} + \left(\sum_{i=1}^{n} |y_i|^p\right)^{1/p}, \quad p \in \mathbb{R}_{\geq 1}.$$

There are also analogous Hölder and Minkowski inequalities for (the Riemann and Lebesgue) integrals.

1.2 Binomial and Generalized Binomial Theorems

The number of ways that $n \in \mathbb{N}$ distinguishable objects can be ordered is given by

$$n\,(n-1)\,(n-2)\ldots 2 \cdot 1 =: n!, \quad 0! := 1,$$

pronounced "n factorial". The number of ways that k objects can be chosen from n, $0 \leq k \leq n$, when order is relevant, is

$$n\,(n-1)\ldots(n-k+1) =: n_{[k]} = \frac{n!}{(n-k)!}, \tag{1.12}$$

which is referred to as the *falling, or descending factorial*.[2]

If the order of the k objects is irrelevant, then $n_{[k]}$ is adjusted by dividing by $k!$, the number of ways of arranging the k chosen objects. Thus, the total number of ways is

$$\frac{n\,(n-1)\cdots(n-k+1)}{k!} = \frac{n!}{(n-k)!\,k!} =: \binom{n}{k}, \quad \binom{n}{0} = 1, \tag{1.13}$$

which is pronounced "n choose k" and referred to as a *binomial coefficient* for reasons which will become clear below. Notice that, both algebraically and intuitively,

$$\binom{n}{k} = \binom{n}{n-k}. \tag{1.14}$$

Example 1.3 *For k even, let $A(k) = 2 \cdot 4 \cdot 6 \cdot 8 \cdot \cdots \cdot k$. Then*

$$A(k) = (1 \cdot 2)\,(2 \cdot 2)\,(3 \cdot 2)\,(4 \cdot 2) \cdots \left(\frac{k}{2} \cdot 2\right) = 2^{k/2} \left(\frac{k}{2}\right)!.$$

[2]Similarly, we denote the *rising, or ascending factorial*, by $n^{[k]} = n\,(n+1)\ldots(n+k-1)$. There are other notational conventions for expressing the falling factorial; for example, William Feller's influential volume I (first edition, 1950, p. 28) advocates $(n)_k$, while Norman L. Johnson (1975) and the references therein (Johnson being the author and editor of numerous important statistical encyclopediae) give reasons for supporting $n^{(k)}$ for the falling factorial (and $n^{[k]}$ for the rising). One still sees the rising factorial denoted by $(n)_k$, which is referred to as the *Pochhammer symbol*, after Leo August Pochhammer, 1841–1920. It will be made clear from context what is meant, so there will be no notational confusion.

With m odd and $C(m) = 1 \cdot 3 \cdot 5 \cdot 7 \cdot \cdots \cdot m$,

$$C(m) = \frac{(m+1)!}{(m+1)(m-1)(m-3)\cdots 6 \cdot 4 \cdot 2} = \frac{(m+1)!}{A(m+1)} = \frac{(m+1)!}{2^{(m+1)/2}\left(\frac{m+1}{2}\right)!}.$$

Thus,

$$C(2i-1) = 1 \cdot 3 \cdot 5 \cdots (2i-1) = \frac{(2i)!}{2^i i!}, \qquad i \in \mathbb{N}, \tag{1.15}$$

a simple result that we will use below. ■

A very useful identity is

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}, \quad k < n, \tag{1.16}$$

which follows because

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

Example 1.4 *Consider the sum $S_n = \sum_{k=0}^{n} \binom{n}{k}$ for $n \in \mathbb{N}$. Imagine that the objects under consideration are the bits in computer memory; they can each take on the value 0 or 1. Among n bits, observe that there are 2^n possible signals that can be constructed. But this is what S_n also gives, because, for a given k, $\binom{n}{k}$ is the number of ways of choosing which of the n bits are set to one, and which are set to zero, and we sum this up over all possible k (0 to n) so that it gives all the possible signals that n binary bits can construct. (That $S_n = 2^n$ also follows directly from the binomial theorem, which is discussed below.) To prove the result via induction, assume it holds for $n-1$, so that, from (1.16),*

$$\sum_{k=0}^{n} \binom{n}{k} = \sum_{k=0}^{n} \left[\binom{n-1}{k} + \binom{n-1}{k-1}\right].$$

Using the fact that $\binom{m}{i} = 0$ for $i > m$,

$$\sum_{k=0}^{n} \binom{n}{k} = \sum_{k=0}^{n-1} \binom{n-1}{k} + \sum_{k=1}^{n} \binom{n-1}{k-1} = 2^{n-1} + \sum_{k=1}^{n} \binom{n-1}{k-1},$$

and, with $j = k-1$, the latter term is

$$\sum_{k=1}^{n} \binom{n-1}{k-1} = \sum_{j=0}^{n-1} \binom{n-1}{j} = 2^{n-1},$$

so that $\sum_{k=0}^{n} \binom{n}{k} = 2^{n-1} + 2^{n-1} = 2(2^{n-1}) = 2^n$. ■

Example 1.5 *To prove the identity*

$$\frac{1}{2} = \sum_{i=0}^{n-1} \binom{n+i-1}{i} \left(\frac{1}{2}\right)^{n+i} =: P_n, \quad n \in \mathbb{N}, \tag{1.17}$$

first note that $P_1 = 1/2$ *and assume* $P_n = 1/2$. *Then,*

$$\begin{aligned}
2P_{n+1} &= \sum_{i=0}^{n} \binom{n+i}{i} \left(\frac{1}{2}\right)^{n+i} \\
&= \sum_{i=0}^{n} \binom{n+i-1}{i} \left(\frac{1}{2}\right)^{n+i} + \sum_{i=0}^{n} \binom{n+i-1}{i-1} \left(\frac{1}{2}\right)^{n+i} \\
&= \underbrace{\sum_{i=0}^{n-1} \binom{n+i-1}{i} \left(\frac{1}{2}\right)^{n+i}}_{=P_n=\frac{1}{2}} + \binom{2n-1}{n} \left(\frac{1}{2}\right)^{2n} + \sum_{i=1}^{n} \binom{n+i-1}{i-1} \left(\frac{1}{2}\right)^{n+i} \\
&\overset{j=i-1}{=} \frac{1}{2} + \binom{2n-1}{n} \left(\frac{1}{2}\right)^{2n} + \sum_{j=0}^{n-1} \binom{n+j}{j} \left(\frac{1}{2}\right)^{n+j+1} \\
&= \frac{1}{2} + \binom{2n-1}{n} \left(\frac{1}{2}\right)^{2n} + P_{n+1} - \binom{2n}{n} \left(\frac{1}{2}\right)^{2n+1}.
\end{aligned}$$

Now note that

$$\begin{aligned}
\binom{2n-1}{n} \left(\frac{1}{2}\right)^{2n} &= \frac{(2n-1)(2n-2)\cdots n}{n!} \left(\frac{1}{2}\right)^{2n} \\
&= \frac{n}{2n} \frac{2n(2n-1)(2n-2)\cdots(n+1)}{n!} \left(\frac{1}{2}\right)^{2n} = \binom{2n}{n} \left(\frac{1}{2}\right)^{2n+1},
\end{aligned}$$

or

$$2P_{n+1} = \frac{1}{2} + P_{n+1} \Leftrightarrow P_{n+1} = \frac{1}{2}.$$

We will use this result in the next example; and prove (1.17) *in a different way, in Example 1.22 below.* ■

Example 1.6 *Prove, for* $N \in \mathbb{N}$,

$$1 = \sum_{k=0}^{N} \binom{2N-k}{N} \left(\frac{1}{2}\right)^{2N-k}. \tag{1.18}$$

This is equivalent to (substitute $i = N-k$, *so* $k = N-i$, *and* $2N-k = 2N-(N-i) = N+i$*)*

$$2^{2N} = \sum_{k=0}^{N} \binom{2N-k}{N} 2^k = \sum_{i=0}^{N} \binom{N+i}{N} 2^{N-i} = 2^N \sum_{i=0}^{N} \binom{N+i}{N} \left(\frac{1}{2}\right)^i,$$

or

$$2^N = \sum_{i=0}^{N} \binom{N+i}{N} \left(\frac{1}{2}\right)^i.$$

But this holds, because, from (1.17), *it follows that*

$$2^{n-1} = \sum_{i=0}^{n-1} \binom{n+i-1}{i} \left(\frac{1}{2}\right)^i,$$

and taking $N = n-1$. *We will use* (1.18) *to prove* (6.22) *is a valid pmf.* ■

By applying (1.16) recursively,

$$\begin{aligned}
\left(\begin{array}{c} n \\ k \end{array} \right) &= \left(\begin{array}{c} n-1 \\ k \end{array} \right) + \left(\begin{array}{c} n-1 \\ k-1 \end{array} \right) \\
&= \left(\begin{array}{c} n-1 \\ k \end{array} \right) + \left(\begin{array}{c} n-2 \\ k-1 \end{array} \right) + \left(\begin{array}{c} n-2 \\ k-2 \end{array} \right) \\
&= \left(\begin{array}{c} n-1 \\ k \end{array} \right) + \left(\begin{array}{c} n-2 \\ k-1 \end{array} \right) + \left(\begin{array}{c} n-3 \\ k-2 \end{array} \right) + \left(\begin{array}{c} n-3 \\ k-3 \end{array} \right) \\
&\vdots \\
&= \sum_{i=0}^{k} \left(\begin{array}{c} n-i-1 \\ k-i \end{array} \right), \quad k < n,
\end{aligned}$$

i.e.,

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \sum_{i=0}^{k} \left(\begin{array}{c} n-i-1 \\ k-i \end{array} \right), \quad k < n. \tag{1.19}$$

In (1.19), replace n with $n+r$, set $k = n$ and rearrange to get

$$\left(\begin{array}{c} n+r \\ n \end{array} \right) = \sum_{i=0}^{n} \left(\begin{array}{c} n+r-i-1 \\ n-i \end{array} \right) = \sum_{i=0}^{n} \left(\begin{array}{c} i+r-1 \\ i \end{array} \right), \tag{1.20}$$

which we will require in §6.3.

Theorem (Binomial Theorem): The relation

$$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i} \tag{1.21}$$

is simple, yet fundamental result that arises in numerous applications. Examples include

$$(x+(-y))^2 = x^2 - 2xy + y^2, \qquad (x+y)^3 = x^3 + 3x^2y + 3xy^2 + y^3,$$

$$0 = (1-1)^n = \sum_{i=0}^{n} \binom{n}{i} (-1)^{n-i}, \qquad 2^n = (1+1)^n = \sum_{i=0}^{n} \binom{n}{i}.$$

Proof: We use induction. Observe first that (1.21) holds for $n = 1$. Then, assuming it holds for $n-1$,

$$\begin{aligned}
(x+y)^n &= (x+y)\,(x+y)^{n-1} = (x+y) \sum_{i=0}^{(n-1)} \binom{(n-1)}{i} x^i y^{(n-1)-i} \\
&= \sum_{i=0}^{n-1} \binom{n-1}{i} x^{i+1} y^{n-(i+1)} + \sum_{i=0}^{n-1} \binom{n-1}{i} x^i y^{n-1-i+1}.
\end{aligned}$$

Then, with $j = i + 1$,

$$\begin{aligned}(x+y)^n &= \sum_{j=1}^{n}\binom{n-1}{j-1}x^j y^{n-j} + \sum_{i=0}^{n-1}\binom{n-1}{i}x^i y^{n-i} \\ &= x^n + \sum_{j=1}^{n-1}\binom{n-1}{j-1}x^j y^{n-j} + \sum_{i=1}^{n-1}\binom{n-1}{i}x^i y^{n-i} + y^n \\ &= x^n + \sum_{i=1}^{n-1}\left\{\binom{n-1}{i-1} + \binom{n-1}{i}\right\}x^i y^{n-i} + y^n \\ &= x^n + \sum_{i=1}^{n-1}\binom{n}{i}x^i y^{n-i} + y^n = \sum_{i=0}^{n}\binom{n}{i}x^i y^{n-i}\end{aligned}$$

proving the theorem.

The binomial theorem can be used for proving the following result, which, in turn, will be used for proving (2.82) below.

Theorem: Let $a \in \mathbb{R}_{>1}$ and $k \in \mathbb{N}$. Then

$$\lim_{n\to\infty} a^n/n^k = \infty. \tag{1.22}$$

Proof: As in Lang, Undergraduate analysis, 2nd ed., 1997, p. 55: Write $a = 1 + b$, so

$$(1+b)^n = 1 + nb + \cdots + \frac{n(n-1)\cdots(n-k)}{(k+1)!}b^{k+1} + \cdots .$$

All the terms in this expansion are positive. The coefficient of b^{k+1} can be written in the form

$$\frac{n^{k+1}}{(k+1)!} + \text{ terms with lower powers of } n.$$

For example, with $k = 3$,

$$\frac{n\,(n-1)\,(n-2)\,(n-3)}{(3+1)!} = \frac{n^{3+1}}{(3+1)!} + \left(-\frac{1}{4}n + \frac{11}{24}n^2 - \frac{1}{4}n^3\right).$$

Hence,

$$\frac{(1+b)^n}{n^k} \geqq \frac{n}{(k+1)!}\left(1 + \frac{c_1}{n} + \cdots + \frac{c_{k+1}}{n^{k+1}}\right)b^{k+1},$$

where $c_1, \ldots, c_{k+1}$ are numbers depending only on k but not on n. Hence when $n \to \infty$, it follows that the expression on the right also $\to \infty$, by the rule for the limit of a product with one factor $n/(k+1)! \to \infty$, while the other factor has the limit b^{k+1} as $n \to \infty$.

Example 1.7 *Let f and g denote functions whose nth derivatives exist. Then, by using the usual product rule for differentiation and an induction argument, we can show that*

$$[fg]^{(n)} = \sum_{j=0}^{n}\binom{n}{j}f^{(j)}g^{(n-j)}, \tag{1.23}$$

where $f^{(j)}$ denotes the jth derivative of f. This is sometimes (also) referred to as Leibniz' rule. This is not the binomial theorem per se*, though it has an obvious association.* ■

Example 1.8 *Consider computing* $I = \int_{-1}^{1} (x^2-1)^j \, dx$, *for any* $j \in \mathbb{N}$. *From the binomial theorem and the basic linearity property* (2.134) *of the Riemann integral,*

$$I = \sum_{k=0}^{j} \left(\begin{array}{c} j \\ k \end{array} \right) (-1)^{j-k} \int_{-1}^{1} x^{2k} dx = \sum_{k=0}^{j} \left(\begin{array}{c} j \\ k \end{array} \right) (-1)^{j-k} \frac{2}{2k+1},$$

which is simple to program and compute as a function of j. *In fact, as shown in the next example, integral* I *can also be expressed as*

$$I = \frac{(-1)^j 2^{2j+1}}{\left(\begin{array}{c} 2j \\ j \end{array} \right) (2j+1)}, \tag{1.24}$$

thus implying the charming and non-obvious combinatoric identity

$$\sum_{k=0}^{j} \left(\begin{array}{c} j \\ k \end{array} \right) \frac{(-1)^k}{2k+1} = \frac{2^{2j}}{\left(\begin{array}{c} 2j \\ j \end{array} \right) (2j+1)}, \quad j \in \mathbb{N},$$

as $(-1)^{-k} = (-1)^k$ *and cancelling a 2 and* $(-1)^j$ *from both sides.* ■

Example 1.9 *We wish to show that*

$$\int_{-1}^{1} (x^2-1)^j \, dx = \frac{(-1)^j 2^{2j+1}}{\left(\begin{array}{c} 2j \\ j \end{array} \right) (2j+1)},$$

thus proving identity (1.24). *Using integration by parts (stated and proven below, in* (2.158)*),*

$$\begin{aligned} \int_{-1}^{1} (x^2-1)^j \, dx &= \int_{-1}^{1} (x-1)^j (x+1)^j dx \\ &= \int_{-1}^{1} (x-1)^j d\left(\frac{(x+1)^{j+1}}{j+1} \right) \\ &= \left[\frac{1}{j+1} (x-1)^j (x+1)^{j+1} \right]_{-1}^{1} - \int_{-1}^{1} \frac{j}{j+1} (x-1)^{j-1} (x+1)^{j+1} dx \\ &= (-1) \frac{j}{j+1} \int_{-1}^{1} (x-1)^{j-1} (x+1)^{j+1} dx. \end{aligned}$$

Repeating this,

$$\begin{aligned}
\int_{-1}^{1}\left(x^2-1\right)^j dx &= (-1)\frac{j}{j+1}\int_{-1}^{1}(x-1)^{j-1}d\left(\frac{(x+1)^{j+2}}{j+2}\right)\\
&= -\left[\frac{j}{(j+1)(j+2)}(x-1)^{j-1}(x+1)^{j+2}\right]_{-1}^{1}\\
&+ (-1)^2\int_{-1}^{1}\frac{j(j-1)}{(j+1)(j+2)}(x-1)^{j-2}(x+1)^{j+2}dx\\
&= (-1)^2\frac{j(j-1)}{(j+1)(j+2)}\int_{-1}^{1}(x-1)^{j-2}(x+1)^{j+2}dx\\
&\vdots\\
&= (-1)^j\frac{j!}{(2j)!/j!}\int_{-1}^{1}(x+1)^{2j}dx\\
&= (-1)^j\frac{1}{\binom{2j}{j}}\left[\frac{(x+1)^{2j+1}}{2j+1}\right]_{-1}^{1} = \frac{(-1)^j 2^{2j+1}}{\binom{2j}{j}(2j+1)}.
\end{aligned}$$

■

A generalization of the left hand sides of (1.12) and (1.13) is obtained by relaxing the positive integer constraint on the upper term in the binomial coefficient:

Definition: For $r \in \mathbb{R}$ and $k \in \mathbb{N}$,

$$\binom{r}{k} := \frac{r\,(r-1)\cdots(r-k+1)}{k!}, \quad \binom{r}{0} := 1. \tag{1.25}$$

The calculations clearly still go through, but the result will, in general, be a real number. Notice that r can be negative, and k can exceed r. Listing 1 gives code for computing (1.25).

Theorem: For $n \in \mathbb{N}$,

$$\binom{-n}{k} = (-1)^k\binom{n+k-1}{k}. \tag{1.26}$$

Note that, for $n = 1$, this reduces to $(-1)^k$.

> Proof: From (1.25),
>
> $$\begin{aligned}
> \binom{-n}{k} &= \frac{(-n)\,(-n-1)\cdots(-n-k+1)}{k!} = (-1)^k\,\frac{(n)\,(n+1)\cdots(n+k-1)}{k!}\\
> &= (-1)^k\binom{n+k-1}{k}.
> \end{aligned}$$

This next example gives a useful result using (1.26), but requires using a Taylor series expansion, which we will develop below in §2.5.6.[3]

[3]It is inspired from having seen it in my first class in statistics, using the book by Mood, Graybill, Boes, Introduction to the Theory of Statistics, 3rd ed., 1976, p. 533; the latter author having been my instructor.

```
function c=c(n,k)

if any(n~=round(n)) | any(n<0),  c=cgeneral(n,k); return,  end

vv=find( (n>=k) & (k>=0) );  if length(vv)==0, c=0; return, end
if length(n)==1, nn=n; else nn=n(vv); end
if length(k)==1, kk=k; else kk=k(vv); end

c=zeros(1,max(length(n),length(k)));
t1 = gammaln(nn+1); t2=gammaln(kk+1); t3=gammaln(nn-kk+1);
c(vv)=round( exp ( t1-t2-t3  )   );

function c=cgeneral(nvec,kvec)
% assumes nvec and kvec have equal length and kvec are positive integers.
c=zeros(length(nvec),1);
for i=1:length(nvec)
  n=nvec(i); k=kvec(i);
  p=1; for j=1:k, p=p*(n-j+1); end
  c(i) = p/gamma(k+1);
end
```

Program Listing 1: Computes (1.25) for possible vector values of n and k.

Example 1.10 *Let* $f(x) = (1-x)^t$, $t \in \mathbb{R}$, *and* $|x| < 1$. *With*

$$f'(x) = -t(1-x)^{t-1}, \quad f''(x) = t(t-1)(1-x)^{t-2},$$

and, in general, $f^{(j)}(x) = (-1)^j t_{[j]} (1-x)^{t-j}$, *the Taylor series expansion (2.283) of* $f(x)$ *around zero is given by*

$$(1-x)^t = f(x) = \sum_{j=0}^{\infty} (-1)^j t_{[j]} \frac{x^j}{j!} = \sum_{j=0}^{\infty} \binom{t}{j} (-x)^j, \quad |x| < 1, \tag{1.27}$$

or $(1+x)^t = \sum_{j=0}^{\infty} \binom{t}{j} x^j$, $|x| < 1$. *For* $t = -1$, *(1.27) and (1.26) yield the familiar* $(1-x)^{-1} = \sum_{j=0}^{\infty} x^j$, *while for* $t = -n$, $n \in \mathbb{N}$, *they imply*

$$(1-x)^{-n} = \sum_{j=0}^{\infty} \binom{-n}{j} (-x)^j = \sum_{j=0}^{\infty} \binom{n+j-1}{j} x^j, \quad |x| < 1. \tag{1.28}$$

Taylor's theorem and properties of the gamma function are used to prove the convergence of these expressions. Some references include Protter and Morrey, 1991, pp. 238-9; Hijab, 1997, p. 91; and Stoll, 2001, Thm. 8.8.4. ■

We will use (1.28) below in Example 1.11.

Theorem: For $n \in \mathbb{N}$,

$$\binom{2n}{n} = (-1)^n \, 2^{2n} \binom{-\frac{1}{2}}{n}. \tag{1.29}$$

We will need this for proving (6.25).

Proof: From (1.25), (1.29) follows from

$$\begin{aligned}
\binom{-\frac{1}{2}}{n} &= \frac{\left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right)\cdots\left(-n+\frac{1}{2}\right)}{n!} = \left(-\frac{1}{2}\right)^n \frac{(2n-1)\,(2n-3)\cdots 3\cdot 1}{n!} \\
&= \left(-\frac{1}{2}\right)^n \frac{1}{n!}\frac{(2n)!}{(2n)\,(2n-2)\cdots 4\cdot 2} = (-1)^n \left(\frac{1}{2}\right)^n \frac{1}{n!}\frac{(2n)!}{2^n\, n!} \\
&= \left(\frac{1}{2}\right)^{2n} (-1)^n \binom{2n}{n}.
\end{aligned} \tag{1.30}$$

Numerically checking (always a good idea), for $n = 3$, both sides numerically resolve to $-5/16$; while for $n = 4$, we get, for both sides, $35/128$. Similarly and easier, from (1.25),

$$(-1)^n \binom{-\frac{1}{2}}{n} = \frac{\left(\frac{1}{2}\right)\left(\frac{3}{2}\right)\cdots\left(n-\frac{1}{2}\right)}{n!} = \binom{n-\frac{1}{2}}{n}.$$

Indeed, for $n = 3$, both sides numerically reduce to $5/16$, while for $n = 4$, both sides give $35/128$. Thus, multiplying (1.30) by $(-1)^n$, we also have

$$\binom{n-\frac{1}{2}}{n} = (-1)^n \binom{-\frac{1}{2}}{n} = \left(\frac{1}{2}\right)^{2n} \binom{2n}{n}.$$

Example 1.11 *Consider proving the identity*

$$\sum_{s=0}^{\infty} \frac{m}{m+s} \binom{m+s}{s} (1-\theta)^s = \theta^{-m}, \quad m \in \mathbb{N}, \quad 0 < \theta < 1. \tag{1.31}$$

The result for $m = 1$ is simple: Recall from (1.14), $\binom{1+s}{s} = \binom{1+s}{1} = 1 + s$. Then

$$\sum_{s=0}^{\infty} \frac{1}{1+s} \binom{1+s}{s} (1-\theta)^s = \sum_{s=0}^{\infty} (1-\theta)^s = \theta^{-1}.$$

For the general case, observe that

$$\frac{m}{m+s}\frac{(m+s)!}{m!s!} = \frac{(m+s-1)!}{(m-1)!s!} = \binom{m+s-1}{s} = (-1)^s \binom{-m}{s}$$

from (1.26). Using this and (1.28) implies that (1.31) is

$$\sum_{s=0}^{\infty} \binom{-m}{s} (-(1-\theta))^s = (1-(1-\theta))^{-m} = \theta^{-m}. \qquad \blacksquare$$

A non-obvious extension of the binomial theorem is

$$(x+y)^{[n]} = \sum_{i=0}^{n} \binom{n}{i} x^{[i]} y^{[n-i]}, \quad x^{[n]} := \prod_{j=0}^{n-1} (x+ja), \tag{1.32}$$

for $n = 0, 1, 2, \ldots$, and x, y, a are real numbers. It holds trivially for $n = 0$, and is easy to see for $n = 1$ and $n = 2$, but otherwise appears difficult to verify, and induction gets messy and doesn't (seem to) lead anywhere. Perhaps somewhat surprisingly, the general proof involves calculus; it is proven below in (1.57). Taking $a = -1$ results in a very important special case. Assuming for now the validity of (1.32), we obtain the following. (Paolella, Fundamental Probability, gives three direct proofs of (1.33).)

Theorem (Vandermonde): For $x, y, n \in \mathbb{N}$,

$$\binom{x+y}{n} = \sum_{i=0}^{n} \binom{x}{i} \binom{y}{n-i}. \tag{1.33}$$

We require this for the next example; but more relevantly, it is the justification that the probability mass function of a hypergeometric random variable indeed sums to one.

Proof: With $a = -1$,

$$k^{[n]} = (k)\,(k-1)\,(k-2)\cdots(k-(n-1)) = \frac{k!}{(k-n)!} = \binom{k}{n} n!,$$

so that (1.32) yields, with $k = x + y$,

$$k^{[n]} = \binom{x+y}{n} n! = \sum_{i=0}^{n} \binom{n}{i} \binom{x}{i} i! \binom{y}{n-i} (n-i)! = n! \sum_{i=0}^{n} \binom{x}{i} \binom{y}{n-i}.$$

Cancelling the $n!$ yields the result.

Example 1.12 *We wish to prove*

$$\binom{r_1 + r_2 + y - 1}{y} = \sum_{i=0}^{y} \binom{r_1 + i - 1}{i} \binom{r_2 + y - i - 1}{y - i}, \tag{1.34}$$

which we will invoke below for proving (1.57)*. From* (1.26)*, it follows that*

$$\binom{r_1 + i - 1}{i} = (-1)^i \binom{-r_1}{i} \quad \text{and} \quad \binom{r_2 + y - i - 1}{y - i} = (-1)^{y-i} \binom{-r_2}{y - i},$$

so that the rhs of the desired equation is

$$\begin{aligned} S &= \sum_{i=0}^{y} \binom{-r_1}{i} \binom{-r_2}{y-i} (-1)^y \\ &\overset{(1.33)}{=} (-1)^y \binom{-(r_1 + r_2)}{y} \overset{(1.26)}{=} \binom{r_1 + r_2 + y - 1}{y}. \end{aligned}$$

■

Definition: If a set of n distinct objects is to be divided into k distinct groups, whereby the size of each group is $n_i, i = 1, \ldots, k$, and $\sum_{i=1}^{k} n_i = n$, then the number of possible divisions is given by

$$\binom{n}{n_1, n_2, \ldots, n_k} := \binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n_k}{n_k} = \frac{n!}{n_1! n_2! \cdots n!}.$$

Note how this reduces to the familiar combinatoric when $k = 2$.

Theorem (Multinomial Theorem): For $r, n \in \mathbb{N}$,

$$\left(\sum_{i=1}^{r} x_i\right)^n = \sum_{\mathbf{n}:\, n_\bullet = n,\, n_i \geq 0} \binom{n}{n_1, \ldots, n_r} \prod_{i=1}^{r} x_i^{n_i}, \tag{1.35}$$

where $\mathbf{n}$ denotes the vector $(n_1, \ldots, n_r)$, and $n_\bullet = \sum_{i=1}^{r} n_i$.

In words, the sum is taken over all nonnegative integer solutions to $\sum_{i=1}^{r} n_i = n$, the number of which $\binom{n+r-1}{n}$, obtained using the usual "stars and bars" trick; see, e.g., Paolella, Fundamental Probability, Ch. 2. With $n = 2$, this is just $(\sum_{i=1}^{r} x_i)^2 = \sum_{i=1}^{r}\sum_{j=1}^{r} x_i x_j$.

Example 1.13 *The expression* $(x_1 + x_2 + x_3)^4$ *corresponds to* $r = 3$ *and* $n = 4$, *so that its expansion will have* $\binom{6}{4} = 15$ *terms, starting with*

$$\begin{aligned}(x_1 + x_2 + x_3)^4 &= \binom{4}{4,0,0} x_1^4 x_2^0 x_3^0 + \binom{4}{0,4,0} x_1^0 x_2^4 x_3^0 + \binom{4}{0,0,4} x_1^0 x_2^0 x_3^4 \\ &\quad + \binom{4}{3,1,0} x_1^3 x_2^1 x_3^0 + \binom{4}{3,0,1} x_1^3 x_2^0 x_3^1 + \cdots,\end{aligned}$$

or in full (and obtained fast and reliably from a symbolic computing environment)

$$\begin{aligned}(x_1 + x_2 + x_3)^4 &= x_1^4 + x_2^4 + x_3^4 \\ &\quad + 4x_1^3 x_2 + 4x_1^3 x_3 + 4x_1 x_2^3 + 4x_1 x_3^3 + 4x_2^3 x_3 + 4x_2 x_3^3 \\ &\quad + 6x_1^2 x_2^2 + 6x_1^2 x_3^2 + 6x_2^2 x_3^2 \\ &\quad + 12x_1^2 x_2 x_3 + 12x_1 x_2^2 x_3 + 12x_1 x_2 x_3^2, \qquad \blacksquare\end{aligned}$$

The proof of (1.35) is by induction on r. For $r = 1$, the theorem clearly holds. Assuming it holds for $r = k$, observe that, with $S = \sum_{i=1}^{k} x_i$,

$$\begin{aligned}\left(\sum_{i=1}^{k+1} x_i\right)^n &= (S + x_{k+1})^n = \sum_{i=0}^{n} \frac{n!}{i!\,(n-i)!} x_{k+1}^i S^{n-i} \\ &= \sum_{i=0}^{n} \frac{n!}{i!\,(n-i)!} \sum_{n_1+\cdots+n_k=n-i} \frac{(n-i)!}{n_1! \cdots n_k!} \prod_{i=1}^{k} x_i^{n_i}\, x_{k+1}^i,\end{aligned}$$

from the binomial theorem and (1.35) for $r = k$. By setting $n_{k+1} = i$, this becomes

$$\left(\sum_{i=1}^{k+1} x_i\right)^n = \sum_{n_1+\cdots+n_{k+1}=n} \frac{n!}{n_1! \cdots n_{k+1}!} \prod_{i=1}^{k} x_i^{n_i}\, x_{k+1}^i,$$

as the sum $\sum_{i=0}^{n}\sum_{n_1+\cdots+n_k=n-i}$ is equivalent to $\sum_{n_1+\cdots+n_{k+1}=n}$ for nonnegative n_i. This is precisely (1.35) for $r = k + 1$, proving the theorem.

Example 1.14 *Using the power series expression of the exponential function*

$$f(x) = \exp(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots, \tag{1.36}$$

we wish to attempt to show that $[\exp(x)]^n = \exp(nx)$. *Applying the multinomial theorem* (1.35) *to* (1.36), *and using* (2.25), *i.e., that* $[f(x)]^n$ *is a composition of two continuous functions, thus allowing the passing of the limit, gives*

$$[f(x)]^n = \left(\lim_{r\to\infty} \sum_{s=0}^{r} \frac{x^s}{s!}\right)^n = \lim_{r\to\infty} \left(\sum_{s=0}^{r} \frac{x^s}{s!}\right)^n. \tag{1.37}$$

Consider first the case with $r = 2$. *Because the sum over* s *starts with a zero instead of a one, we use the terms* $n_0, n_1, \ldots,$ *instead of starting with* n_1. *We begin by explicitly isolating the terms that give rise to* x^0, x^1, *and* x^2 *in the expansion. In particular, as* $1^{n_0} = 1$ *for* $n_0 \geq 1$, *we seek the terms of the form* x^{n_1} *and* $(x^2)^{n_2}$ *such that we get products of the form* x^0, x^1, *and* x^2. *We let* k *indicate the power of* x *of this product, so the term with* $k = 0$ *will be* $1^{n_0}x^{n_1}(x^2)^{n_2}$ *with* $n_0 = n$, $n_1 = n_2 = 0$; *the* $k = 1$ *term will be* $1^{n_0}x^{n_1}(x^2)^{n_2}$ *with* $n_0 = n-1$, $n_1 = 1$, $n_2 = 0$; *and finally, for* $k = 2$, $1^{n_0}x^{n_1}(x^2)^{n_2}$ *with either* $n_0 = n-1$, $n_1 = 0$, $n_2 = 1$, *or* $n_0 = n-2$, $n_1 = 2$, $n_2 = 0$. *The expansion, showing explicitly the* $k = 0, 1, 2$ *cases, is*

$$\begin{aligned}
\left(1 + x + \frac{x^2}{2}\right)^n &= \sum_{\mathbf{n}:\, n_\bullet = n,\, n_i \geq 0} \binom{n}{n_0, n_1, n_2} 1^{n_0} \left(\frac{x}{1!}\right)^{n_1} \left(\frac{x^2}{2!}\right)^{n_2} \\
&= \binom{n}{n,0,0} 1^n \left(\frac{x}{1!}\right)^0 \left(\frac{x^2}{2!}\right)^0 + \binom{n}{n-1,1,0} 1^{n-1} \left(\frac{x}{1!}\right)^1 \left(\frac{x^2}{2!}\right)^0 \\
&\quad + \binom{n}{n-1,0,1} 1^{n-1} \left(\frac{x}{1!}\right)^0 \left(\frac{x^2}{2!}\right)^1 + \binom{n}{n-2,2,0} 1^{n-2} \left(\frac{x}{1!}\right)^2 \left(\frac{x^2}{2!}\right)^0 \\
&\quad + \sum_{n_1+2n_2=3} \binom{n}{n_0, n_1, n_2} 1^{n_0} \left(\frac{x}{1!}\right)^{n_1} \left(\frac{x^2}{2!}\right)^{n_2} + \cdots.
\end{aligned}$$

The last line gathers terms that give rise to powers of x *of* $k = 3$, *i.e., we require, as always,* $n_0 + n_1 + n_2 = n$, *and also* $n_1 + 2n_2 = 3$. *The remaining terms not shown are for* $k = 4$, $k = 5$, *etc.. Simplifying, we get*

$$\begin{aligned}
\left(1 + x + \frac{x^2}{2}\right)^n &= 1 + \frac{n!}{(n-1)!}x + \frac{n!}{(n-1)!}\frac{x^2}{2} + \frac{n!}{(n-2)!2!}x^2 + \cdots + C_k x^k + \cdots \\
&= 1 + nx + \frac{nx^2}{2} + \frac{n(n-1)}{2}x^2 + \cdots \\
&= 1 + (nx) + \frac{(nx)^2}{2} + \cdots,
\end{aligned}$$

which indeed begins to look like $f(nx)$ *from* (1.36).

Of course, more work is required to actually determine that

$$C_k = \sum_{n_1+2n_2+\cdots rn_r=k} \binom{n}{n_0, \ldots, n_r} \prod_{i=1}^{r} (i!)^{-n_i} = \frac{n^k}{k!}.$$

This might be amenable to induction; we do not attempt it here. For $k = 3$, *the solutions to* $n_1 + 2n_2 + \cdots rn_r = 3$ *(with* n_0 *such that* $\sum_{i=0}^{r} n_r = n$*) are* $(n_0, \ldots, n_r) = (n_0, 3, 0, \ldots, 0)$,

$(n_0, 1, 1, 0, \ldots, 0)$, *and* $(n_0, 0, 0, 1, 0, \ldots, 0)$. *Thus,*

$$C_3 = \frac{n!}{(n-3)!3!}\frac{1}{1^3} + \frac{n!}{(n-2)!}\frac{1}{1^1}\frac{1}{2^1} + \frac{n!}{(n-1)!}\frac{1}{6^1} = \frac{n^3}{3!},$$

suggesting this method of proof is viable. ■

1.3 Gamma and Beta Functions

> Young people today love luxury. They have bad manners, despise authority, have no respect for older people, and chatter when they should be working.
>
> (Socrates, 470–399 BC)

There exist (infinitely many) elementary functions f such that there is no elementary function $F(x)$ satisfying $F'(x) = f(x)$. Important (because of their application to real problems) examples of such functions f include the gamma function discussed here, the beta function discussed below, and the Gauss error function $\exp(x^2)$. Perhaps surprisingly, it is true also for e^x/x and $1/\ln x$.

The gamma function can be expressed as

$$\Gamma(x) := \int_0^\infty t^{x-1}e^{-t}dt, \quad x \in \mathbb{R}_{>0}. \tag{1.38}$$

Being an improper integral, we need to confirm its existence. This requires use of the comparison test for improper integrals, given in (2.197):

> Proof: In view of the inequality $t^{x-1}e^{-t} \leq t^{x-1}$ for $t > 0$, the existence of the integral $\int_0^1 t^{x-1}e^{-t}dt$ follows from that of $\int_0^1 t^{x-1}dt$, provided that $x > 0$. Next, since $t^{x+1}e^{-t} \to 0$ as $t \to +\infty$, we have, for some $H > 0$, $t^{x-1}e^{-t} \leq Ht^{-2}$ for $t \geq 1$. Hence, the existence of $\int_1^\infty t^{x-1}e^{-t}dt$ follows from that of $\int_1^\infty t^{-2}dt$.
>
> The convergence of the improper integral from 1 to infinity is also addressed in Example 2.56.

As mentioned above, an expression for $\Gamma(x)$ in terms of "elementary" functions does not exist for general x. However, a basic integration by parts (see (2.158)) shows that

$$\Gamma(x) = (x-1)\Gamma(x-1), \quad x \in \mathbb{R}_{>1}. \tag{1.39}$$

Thus, for $n \in \mathbb{N}$,

$$\Gamma(n) = (n-1)!. \tag{1.40}$$

As in Andrews, Askey and Roy (1999, pp. 2-3; see also p. 35), suppose that $x \geq 0$ and $n \geq 0$ are integers. Write

$$x! = \frac{(x+n)!}{(x+1)_n}, \tag{1.41}$$

where $(a)_n$ denotes the rising factorial (using Pochhammer's notation; see above)

$$(a)_n = a(a+1)\cdots(a+n-1) \text{ for } n > 0, \quad (a)_0 = 1,$$

and a is any real number. Rewrite (1.41) as

$$x! = \frac{n!(n+1)_x}{(x+1)_n} = \frac{n!n^x}{(x+1)_n} \cdot \frac{(n+1)_x}{n^x}.$$

Since

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

we conclude that

$$x!=\lim_{n\to\infty}\frac{n!n^x}{(x+1)_n}.$$

This, along with (1.40), is (also) used to *define* the gamma function as

$$\Gamma(x)=\lim_{n\to\infty}\frac{n!n^x}{x(x+1)\cdots(x+n)},\quad x>0, \tag{1.42}$$

known as the Gauss product formula. The equivalence of (1.38) and (1.42) is proven in Appendix 6.1. We will require (1.42) below in Example 6.3. Appendix §6.1 contains further results on the gamma (and beta) functions. We will also require the identity (we will need it directly below, and for deriving the important relationship (1.51) between the gamma and beta functions)

$$\Gamma(a)=2\int_0^\infty u^{2a-1}e^{-u^2}du. \tag{1.43}$$

This follows directly by using the substitution $u=x^{1/2}$ in (1.38) (recall x is positive), so that $x=u^2$ and $dx=2udu$. Another useful fact that follows from (1.43) and Example 5.20 (and letting, say, $v=\sqrt{2}u$) is that

$$\Gamma(1/2)=\sqrt{\pi}, \tag{1.44}$$

which we will use often.

Example 1.15 *Recall the Gaussian probability density function. For $Z\sim \mathrm{N}(0,1)$, we wish to compute the even positive moments, $\mathbb{E}[Z^{2r}]$, for $r\in\mathbb{N}$. With $u=z^2/2$, $z=(2u)^{1/2}$ (because z is positive), and $dz=(2u)^{-1/2}du$,*

$$\begin{aligned}\mathbb{E}\left[Z^{2r}\right]&=\int_{-\infty}^{\infty}z^{2r}f_Z(z)\,dz=\frac{2}{\sqrt{2\pi}}\int_0^\infty z^{2r}e^{-\frac{1}{2}z^2}dz\\&=\frac{2^{r+1-1/2}}{\sqrt{2\pi}}\int_0^\infty u^{r-1/2}e^{-u}du=\frac{2^r\Gamma\left(r+\frac{1}{2}\right)}{\sqrt{\pi}}.\end{aligned}$$

That is, for $s=2r$ and recalling that $\Gamma(a+1)=a\Gamma(a)$ and $\Gamma(1/2)=\sqrt{\pi}$,

$$\mathbb{E}[Z^s]=\frac{1}{\sqrt{\pi}}2^{s/2}\Gamma\left(\frac{1}{2}(1+s)\right)=(s-1)(s-3)(s-5)\cdot\cdots\cdot 3\cdot 1. \tag{1.45}$$

This can also be written

$$\mathbb{E}[Z^s]=\mathbb{E}\left[Z^{2r}\right]=\frac{(2r)!}{2^r r!}, \tag{1.46}$$

which follows because (in the numerator, note $(2r)!=(2r)(2r-1)(2r-2)!$)

$$M(r):=\frac{(2r)!}{2^r r!}=\left[\frac{(2r-1)\,2r}{2r}\right]\frac{(2(r-1))!}{2^{r-1}(r-1)!}=(2r-1)\,M(r-1),$$

e.g.,

$$\mathbb{E}[Z^6]=\mathbb{E}[Z^{2\cdot 3}]=M(3)=5\cdot M(2)=5\cdot 3\cdot M(1)=5\cdot 3\cdot 1.$$

With $X = \sigma Z + \mu \sim \mathrm{N}(\mu, \sigma^2)$,

$$\mathbb{E}\left[(X-\mu)^{2r}\right] = \sigma^{2r}\mathbb{E}\left[Z^{2r}\right] = \left(2\sigma^2\right)^r \pi^{-1/2}\Gamma\left(r+\frac{1}{2}\right). \tag{1.47}$$

(The reader should directly check that (1.47) reduces to $3\sigma^4$ *for* $r = 2$*.) An expression for the even raw moments of* X *can be obtained via (1.46) and the binomial formula applied to* $(\sigma Z + \mu)^{2r}$*.*

For odd moments, similar calculations give[4]

$$\begin{aligned}\int_{-\infty}^{\infty} z^{2r+1} f_Z(z)\,dz &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} z^{2r+1} e^{-\frac{1}{2}z^2} dz + \frac{1}{\sqrt{2\pi}}\int_{0}^{\infty} z^{2r+1} e^{-\frac{1}{2}z^2} dz \\ &= -\frac{2^r\Gamma(r+1)}{\sqrt{2\pi}} + \frac{2^r\Gamma(r+1)}{\sqrt{2\pi}} = 0.\end{aligned}$$

Thus, for example, the skewness and kurtosis of $X = \sigma Z + \mu$ *is zero and three, respectively, recalling that those measures are location and scale invariant.*

To calculate $\mathbb{E}|Z| := \mathbb{E}\big[\,|Z|\,\big]$*, use the same* u *substitution as above to give*

$$\begin{aligned}\mathbb{E}|Z| &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} |z|\, f_Z(z)\,dz = \frac{2}{\sqrt{2\pi}}\int_{0}^{\infty} z e^{-\frac{1}{2}z^2} dz \\ &= \frac{2}{\sqrt{2\pi}}\int_{0}^{\infty} (2u)^{1/2} e^{-u} (2u)^{-1/2}\,du = \sqrt{\frac{2}{\pi}},\end{aligned} \tag{1.48}$$

where $\int_0^\infty e^{-u}du = 1$*.* ■

The beta function is an integral expression of two parameters, denoted $B(\cdot,\cdot)$ and defined to be

$$B(a,b) := \int_0^1 x^{a-1}(1-x)^{b-1}\,dx, \quad a,b \in \mathbb{R}_{>0}. \tag{1.49}$$

By substituting $x = \sin^2\theta$ into (1.49) we obtain (and as directly used below) that

$$B(a,b) = \int_0^{\pi/2} \left(\sin^2\theta\right)^{a-1}\left(\cos^2\theta\right)^{b-1} 2\sin\theta\cos\theta d\theta = 2\int_0^{\pi/2} (\sin\theta)^{2a-1}(\cos\theta)^{2b-1}\,d\theta. \tag{1.50}$$

Closed-form expressions do not exist for general a and b; however, the identity

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{1.51}$$

can be used for its evaluation in terms of the gamma function. There are several ways of proving this. Here is one. Using polar coordinates $x = r\cos\theta$, $y = r\sin\theta$, $dx\,dy = r\,dr\,d\theta$

[4]It should be clear from the symmetry of f_Z that $\int_{-k}^{k} z^{2r+1} f_Z(z)\,dz = 0$ for any $k > 0$ (see §2.4.3). Recall also from §2.4.3 that, for a general density f_X, in order to claim that $\int_{-\infty}^{\infty} x^{2r+1} f_X(x)\,dx = 0$, both $\lim_{k\to\infty}\int_0^k x^{2r+1} f_X(x)\,dx$ and $-\lim_{k\to\infty}\int_{-k}^0 x^{2r+1} f_X(x)\,dx$ must converge to the same finite value.

(see (5.42) in §5.6) along with (1.43) and (1.50),

$$\begin{aligned}\Gamma(a)\Gamma(b) &= 4\int_0^\infty\int_0^\infty x^{2a-1}y^{2b-1}e^{-\left(x^2+y^2\right)}dxdy \\ &= 4\int_0^{2\pi}\int_0^\infty r^{2(a+b)-2+1}e^{-r^2}(\cos\theta)^{2a-1}(\sin\theta)^{2b-1}drd\theta \\ &= 4\left(\int_0^\infty r^{2(a+b)-1}e^{-r^2}dr\right)\left(\int_0^{2\pi}(\cos\theta)^{2a-1}(\sin\theta)^{2b-1}d\theta\right) \\ &= \Gamma(a+b)B(a,b).\end{aligned}$$

A direct proof without the use of polar coordinates can be found in Hijab (1997, p. 193). If $a=b$, then, from symmetry (or use the substitution $y=1-x$) and use of (1.51), it follows that

$$\int_0^{1/2}x^{a-1}(1-x)^{a-1}dx=\int_{1/2}^1 x^{a-1}(1-x)^{a-1}dx=\frac{1}{2}\frac{\Gamma^2(a)}{\Gamma(2a)}, \tag{1.52}$$

where $\Gamma^2(a)$ is just a shorthand notation for $[\Gamma(a)]^2$. We will use this result now:

Theorem (Legendre's duplication formula):

$$\Gamma(2a)=\frac{2^{2a-1}}{\sqrt{\pi}}\Gamma(a)\Gamma\left(a+\frac{1}{2}\right). \tag{1.53}$$

Proof: Use (1.52) with $u=4x(1-x)$ (and, as $0\le x\le 1/2$, $x=\left(1-\sqrt{1-u}\right)/2$ and $dx=1/(4\sqrt{1-u})du$) to get

$$\begin{aligned}\frac{\Gamma^2(a)}{\Gamma(2a)} &= 2\int_0^{1/2}x^{a-1}(1-x)^{a-1}dx=\frac{2}{4^{a-1}}\int_0^{1/2}(4x(1-x))^{a-1}dx \\ &= \frac{2}{4^{a-1}}\int_0^1 u^{a-1}\frac{1}{4}(1-u)^{-1/2}du=2^{1-2a}\frac{\Gamma(a)\Gamma(1/2)}{\Gamma(a+1/2)}.\end{aligned}$$

As $\Gamma(1/2)=\sqrt{\pi}$, the result follows.

Example 1.16 *From Legendre's duplication formula (1.53) with $i\in\mathbb{N}$ and using (1.15), we obtain (but also note that the result follows directly from (1.40) and (1.44))*

$$\begin{aligned}\Gamma\left(i+\frac{1}{2}\right) &= \frac{\sqrt{\pi}\Gamma(2i)}{2^{2i-1}\Gamma(i)}=\frac{\sqrt{\pi}}{2^{2i-1}}\frac{(2i-1)!}{(i-1)!}=\frac{\sqrt{\pi}}{2^{2i-1}}\frac{i}{2i}\frac{(2i)!}{i!} \\ &= \frac{\sqrt{\pi}}{2^{2i}}\frac{(2i)!}{i!}=\frac{\sqrt{\pi}}{2^{2i}}2^iC(2i-1) \\ &= \frac{1\cdot 3\cdot 5\cdots(2i-1)}{2^i}\sqrt{\pi},\end{aligned} \tag{1.54}$$

which is required, for example, when deriving properties of the noncentral Student's t and related distributions; see, e.g., Paolella, Intermediate Probability. ■

Example 1.17 *To express $\int_0^1\sqrt{1-x^4}\,dx$ in terms of the beta function, let $u=x^4$ and $dx=(1/4)u^{1/4-1}\,du$, so that*

$$\int_0^1\sqrt{1-x^4}\,dx=\frac{1}{4}\int_0^1 u^{-3/4}(1-u)^{1/2}du=\frac{1}{4}B\left(\frac{1}{4},\frac{3}{2}\right). \qquad \blacksquare$$

Example 1.18 *To compute*

$$I = \int_0^s x^a (s-x)^b \, dx, \qquad s \in (0,1), \quad a, b > 0,$$

use $u = 1 - x/s$ *(so that* $x = (1-u)\, s$ *and* $dx = -s du$*) to get*

$$\begin{aligned} I &= \int_0^s x^a (s-x)^b \, dx = -s \int_1^0 ((1-u)\, s)^a (s - (1-u)\, s)^b \, du \\ &= s^{a+b+1} \int_0^1 (1-u)^a u^b \, du = s^{a+b+1} B(b+1, a+1). \end{aligned}$$ ■

Example 1.19 *To compute*

$$I = \int_{-1}^1 \left(1-x^2\right)^a (1-x)^b \, dx,$$

use $1 - x^2 = (1-x)(1+x)$ *and* $u = (1+x)/2$ *(*$x = 2u - 1$, $dx = 2du$*) to get*

$$I = 2^{2a+b+1} \int_0^1 u^a (1-u)^{a+b} \, du = 2^{2a+b+1} B(a+1, a+b+1).$$ ■

Example 1.20 *The moment generating function (m.g.f.) of a location-zero, scale-one logistic random variable is (with* $y = (1 + \mathrm{e}^{-x})^{-1}$*), for* $|t| < 1$*,*

$$\begin{aligned} \mathbb{M}_X(t) = \mathbb{E}\left[e^{tX}\right] &= \int_{-\infty}^{\infty} \left(e^{-x}\right)^{1-t} \left(1+e^{-x}\right)^{-2} dx \\ &= \int_0^1 \left(\frac{1-y}{y}\right)^{1-t} y^2 y^{-1} (1-y)^{-1} dy = \int_0^1 (1-y)^{-t} y^t dy \\ &= B(1-t, 1+t) = \Gamma(1-t)\Gamma(1+t). \end{aligned}$$

If, in addition, $t \neq 0$*, the m.g.f. can also be expressed as*

$$\mathbb{M}_X(t) = t\Gamma(t)\Gamma(1-t) = t\frac{\pi}{\sin \pi t}, \tag{1.55}$$

where the second identity is called Euler's reflection formula*: Andrews, Askey and Roy (1999, pp. 9-10) provide four different methods for proving Euler's reflection formula; see also Jones (2001, pp. 217-18), Havil (2003, p. 59), Schiff (1999, p. 174), and Duren, Invitation to Classical Analysis, 2012, §9.5. Notice also, from (1.55) with* $t = 1/2$*, it follows that* $\Gamma(1/2) = \sqrt{\pi}$*.* ■

Example 1.21 *An interesting relation both theoretically and computationally is given by*

$$\sum_{j=k}^{n} \binom{n}{j} p^j (1-p)^{n-j} = \frac{\Gamma(n+1)}{\Gamma(k)\,\Gamma(n-k+1)} \int_0^p x^{k-1} (1-x)^{n-k} \, dx, \tag{1.56}$$

for $0 \leq p \leq 1$ *and* $k = 1, 2, \ldots$*, where* $\binom{n}{j}$ *is a binomial coefficient, and can be proven by repeated integration by parts. To motivate this, take* $k = 1$*. From the binomial theorem (1.21)*

with $x = p = 1 - y$, it follows directly that the lhs of (1.56) is $1 - (1-p)^n$. The rhs is, with $y = 1 - x$,

$$\frac{n!}{(n-1)!}\int_0^p (1-x)^{n-1}\,dx = -n\int_1^{1-p} y^{n-1}dy = y^n|_{1-p}^1 = 1-(1-p)^n\,.$$

For $k = 2$, the lhs of (1.56) is easily seen to be $1-(1-p)^n - np\,(1-p)^{n-1}$, while the rhs is, using $y = 1 - x$,

$$\begin{aligned}\frac{n!}{1!\,(n-2)!}\int_0^p x\,(1-x)^{n-2}\,dx &= -n\,(n-1)\int_1^{1-p}(1-y)\,y^{n-2}dy\\ &= n\,(n-1)\left[\frac{y^{n-1}}{n-1}\Big|_{1-p}^1 - \frac{y^n}{n}\Big|_{1-p}^1\right]\\ &= 1-np\,(1-p)^{n-1}-(1-p)^n\,,\end{aligned}$$

after some rearranging. ■

Example 1.22 *By substituting $2n-1$ for n; n for k; and taking $p = 1/2$ in (1.56), we get*

$$\sum_{i=n}^{2n-1}\binom{2n-1}{i}\left(\frac{1}{2}\right)^{2n-1} = \frac{\Gamma(2n)}{\Gamma^2(n)}\int_0^{1/2} u^{n-1}(1-u)^{n-1}\,du = \frac{1}{2}$$

from (1.52), directly showing (1.17) in Example 1.5. ■

Theorem (Binomial Theorem Extension): For $n = 0, 1, 2, \ldots$, and $x, y, a \in \mathbb{R}$, define

$$x^{[n]} := \prod_{j=0}^{n-1}(x + ja).$$

Then

$$(x+y)^{[n]} = \sum_{i=0}^{n}\binom{n}{i}x^{[i]}y^{[n-i]}, \tag{1.57}$$

as was first stated above in (1.32), without proof.

Proof: As

$$\begin{aligned}\prod_{j=0}^{k}(x+ja) &= (x)\,(x+a)\cdots(x+ka)\\ &= a^{k+1}\left(\frac{x}{a}\right)\left(\frac{x}{a}+1\right)\cdots\left(\frac{x}{a}+k\right) = a^{k+1}\frac{\Gamma(k+1+x/a)}{\Gamma(x/a)},\end{aligned}$$

(1.57) can be expressed as the conjecture

$$\begin{aligned}(x+y)^{[n]} &\stackrel{?}{=} \sum_{i=0}^{n}\binom{n}{i}x^{[i]}y^{[n-i]}\\ \prod_{j=0}^{n-1}(x+y+ja) &\stackrel{?}{=} \sum_{i=0}^{n}\binom{n}{i}\left(\prod_{j=0}^{i-1}(x+ja)\right)\left(\prod_{j=0}^{n-i-1}(y+ja)\right)\\ a^n\frac{\Gamma(n+(x+y)/a)}{\Gamma((x+y)/a)} &\stackrel{?}{=} \sum_{i=0}^{n}\binom{n}{i}\left(a^i\frac{\Gamma(i+x/a)}{\Gamma(x/a)}\right)\left(a^{n-i}\frac{\Gamma(n-i+y/a)}{\Gamma(y/a)}\right)\end{aligned}$$

or

$$\frac{\Gamma\left(n+(x+y)/a\right)}{\Gamma\left((x+y)/a\right)} \stackrel{?}{=} \sum_{i=0}^{n}\binom{n}{i}\frac{\Gamma\left(i+x/a\right)}{\Gamma\left(x/a\right)}\frac{\Gamma\left(n-i+y/a\right)}{\Gamma\left(y/a\right)}$$

or

$$\frac{\Gamma\left(x/a\right)\Gamma\left(y/a\right)}{\Gamma\left((x+y)/a\right)} \stackrel{?}{=} \sum_{i=0}^{n}\binom{n}{i}\frac{\Gamma\left(i+x/a\right)\Gamma\left(n-i+y/a\right)}{\Gamma\left(n+(x+y)/a\right)}, \tag{1.58}$$

or, equivalently, using (1.51),

$$B\left(\frac{x}{a},\frac{y}{a}\right) \stackrel{?}{=} \sum_{i=0}^{n}\binom{n}{i}B\left(\frac{x}{a}+i,\frac{y}{a}+n-i\right). \tag{1.59}$$

We now need to prove (1.58) and (1.59). This will be done using results from probability theory. In turn, this proves (1.57).

Let $X_i \overset{\text{ind}}{\sim} \text{Gam}\left(a_i,c\right)$, $i=1,2$, and define $S=X_1+X_2$, which follows a $\text{Gam}\left(a_1+a_2,c\right)$ distribution. The linearity of expectation, and the binomial theorem, imply

$$\mathbb{E}\left[S^k\right]=\mathbb{E}\left[\left(X_1+X_2\right)^k\right]=\sum_{i=0}^{k}\binom{k}{i}\mathbb{E}\left[X_1^i\right]\mathbb{E}\left[X_2^{k-i}\right],$$

or, using that, for $X\sim\text{Gamma}(\alpha,\beta)$, $\mathbb{E}\left[X^k\right]=\frac{\Gamma(k+\alpha)}{\beta^k\Gamma(\alpha)}$ for $k>-\alpha$,

$$\frac{\Gamma\left(k+a_1+a_2\right)}{c^k\Gamma\left(a_1+a_2\right)}=\sum_{i=0}^{k}\binom{k}{i}\frac{\Gamma\left(i+a_1\right)}{c^i\Gamma\left(a_1\right)}\frac{\Gamma\left(k-i+a_2\right)}{c^{k-i}\Gamma\left(a_2\right)}. \tag{1.60}$$

That is, noting the c-terms cancel,

$$\frac{(k+a_1+a_2-1)!}{(a_1+a_2-1)!k!}=\sum_{i=0}^{k}\frac{(i+a_1-1)!}{i!\,(a_1-1)!}\frac{(k-i+a_2-1)!}{(k-i)!\,(a_2-1)!},$$

or

$$\binom{k+a_1+a_2-1}{k}=\sum_{i=0}^{k}\binom{i+a_1-1}{i}\binom{k-i+a_2-1}{k-i},$$

which is precisely (1.34). Rearranging (1.60) gives (1.58), i.e.,

$$\frac{\Gamma\left(a_1\right)\Gamma\left(a_2\right)}{\Gamma\left(a_1+a_2\right)}=\sum_{i=0}^{k}\binom{k}{i}\frac{\Gamma\left(i+a_1\right)\Gamma\left(k-i+a_2\right)}{\Gamma\left(k+a_1+a_2\right)}.$$

Using (1.51), this can be expressed as

$$B\left(a_1,a_2\right)=\sum_{i=0}^{k}\binom{k}{i}B\left(a_1+i,a_2+k-i\right),$$

which gives (1.59). The latter result can also be obtained, faster, by letting $X\sim\text{Beta}\left(a_1,a_2\right)$. In particular, and using the binomial theorem,

$$\begin{aligned}
1 &= \int_0^1 f_X\left(x\right)dx=\frac{1}{B\left(a_1,a_2\right)}\int_0^1\left(x+1-x\right)^k x^{a_1-1}\left(1-x\right)^{a_2-1}dx\\
&= \frac{1}{B\left(a_1,a_2\right)}\sum_{i=0}^{k}\binom{k}{i}\int_0^1 x^{a_1+i-1}\left(1-x\right)^{a_2-1+k-i}dx\\
&= \frac{1}{B\left(a_1,a_2\right)}\sum_{i=0}^{k}\binom{k}{i}B\left(a_1+i,a_2+k-i\right).
\end{aligned}$$

2 Univariate Calculus

> Leibniz never married; he had considered it at the age of fifty; but the person he had in mind asked for time to reflect. This gave Leibniz time to reflect, too, and so he never married. (Bernard Le Bovier Fontenelle)

2.1 Limits and Continuity

A *sequence* is a function $f : \mathbb{N} \to \mathbb{R}$, with $f(n)$, $n \in \mathbb{N}$, being the nth *term* of f. We often denote the sequence of f as $\{s_n\}$, where $s_n = f(n)$. Let $\{s_n\}$ be a sequence.

As repeated in the definition just below: If for any given $\epsilon > 0$, $\exists\, a \in \mathbb{R}$ and $\exists\, N \in \mathbb{N}$ such that, $\forall n \geq N$, $|a - s_n| < \epsilon$, then the sequence is *convergent*, and *converges to* (the unique value) a. If $\{s_n\}$ converges to a, then we write $\lim_{n\to\infty} s_n = a$. If $\{s_n\}$ does not converge, then it is said to *diverge*. Sequence $\{s_n\}$ is *strictly increasing* if $s_{n+1} > s_n$, and *increasing* if $s_{n+1} \geq s_n$. The sequence is *bounded from above* if $\exists c \in \mathbb{R}$ such that $s_n \leq c$ for all n. Similar definitions apply to *decreasing*, *strictly decreasing*, and *bounded from below*. A simple but fundamental result is that, if $\{s_n\}$ is bounded from above and increasing, or bounded from below and decreasing, then it is convergent.

Definition: The sequence $\{a_k\} \in \mathbb{R}$ converges to $a \in \mathbb{R}$ if:

$$\forall \epsilon > 0, \;\; \exists K \in \mathbb{N} \;\; \text{such that} \;\; \forall k > K, \;\; |a_k - a| < \epsilon. \tag{2.1}$$

Point a is the *limit* of $\{a_k\}$ if $\{a_k\}$ converges to a, in which case one writes $\lim_{k\to\infty} a_k = a$. If the limit exists, then it is unique, as shown next.

Theorem: A convergent sequence has a unique limit.

> Proof: Suppose $a_n \to a$ as well as $a_n \to b$. If $b \neq a$, let $\epsilon := |a - b|$. Since $a_n \to a$, there is $n_1 \in \mathbb{N}$ such that $|a_n - a| < \epsilon/2$ for all $n \geq n_1$, and since $a_n \to b$, there is $n_2 \in \mathbb{N}$ such that $|a_n - b| < \epsilon/2$ for all $n \geq n_2$. Let $n_0 := \max\{n_1, n_2\}$. Then
>
> $$|a - b| \leq |a - a_{n_0}| + |a_{n_0} - b| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = |a - b|. \tag{2.2}$$
>
> This contradiction shows that $b = a$.

Theorem (Squeeze Theorem): Suppose $\{a_n\}, \{b_n\}, \{c_n\} \in \mathbb{R}$ are sequences for which there exists $n_o \in \mathbb{N}$ such that $a_n \leq b_n \leq c_n$ for all $n \in \mathbb{N}$, $n \geq n_o$, and that $\lim\limits_{n\to\infty} a_n = \lim\limits_{n\to\infty} c_n = L$. Then

$$\textbf{the sequence } \{b_n\} \textbf{ converges and } \lim_{n\to\infty} b_n = L. \tag{2.3}$$

> Proof: For any $\epsilon > 0$, $\exists n_a \in \mathbb{N}$ such that, $\forall n \geq n_a, a_n \in N_\epsilon(L)$; and $\exists n_c \in \mathbb{N}$ such that, $\forall n \geq n_c, c_n \in N_\epsilon(L)$. Set $N = \max(n_o, n_a, n_c)$, so that $\forall n \geq N, \{a_n, c_n\} \in N_\epsilon(L)$ and $a_n \leq b_n \leq c_n$. Then $b_n \in N_\epsilon(L)$ for $n \geq N$, which is the definition of convergence of sequence $\{b_n\}$.

Theorem: Let $a \in \mathbb{R}$. Then

$$\lim_{n\to\infty} \frac{a^n}{n!} = 0. \tag{2.4}$$

Proof: Choose $m \in \mathbb{N}$ such that $|a| < m$. Then for $n > m$,

$$0 \leq \left|\frac{a^n}{n!}\right| = \frac{|a|^n}{m!}\left(\prod_{j=m+1}^{n}\frac{1}{j}\right) < \frac{|a|^n}{m!}\left(\frac{1}{m^{n-m}}\right) = \frac{m^m}{m!}\left(\frac{|a|}{m}\right)^n.$$

Since m is a constant and $|a| < m$, $(|a|/m)^n \to 0$. The Squeeze Theorem (2.3) implies that $|a^n/n!| \to 0$, and thus, from definition (2.1), $a^n/n! \to 0$.

Theorem: Let $x > 0$. We wish to show

$$0 < \sum_{n=0}^{\infty}\frac{x^{2n}}{(2n)!} < \sum_{n=0}^{\infty}\frac{x^n}{n!}, \tag{2.5}$$

where (not importantly right now) the latter expression is equal to e^x, as will be shown later.

Proof: Inequality (2.5) is obvious for $0 < x \leq 1$. Next, for each fixed $x > 1$, the ratio of the two terms in the sums is $(x^n/n!)/(x^{2n}/(2n)!) = (2n)!/n! \times x^{-n}$. From the reciprocal result of (2.4), the ratio of the two terms from (2.5) goes to infinity as n increases. As the sums are infinite, the inequality must hold. We will use this result in Example 2.89.

Theorem: Let $\{a_n\}$ be a convergent sequence with $a_n \to a \in \mathbb{R}$. Then

$$|a_n| \to |a|. \tag{2.6}$$

Proof: The reverse triangle inequality (1.8) implies $0 \leq ||a_n| - |a|| \leq |a_n - a|$. This holds $\forall n \in \mathbb{N}$, so $|a_n| \to |a|$.

Theorem: Let x_n and y_n be sequences such that $\lim_{n\to\infty} x_n = x$ and $\lim_{n\to\infty} y_n = y$. Then:

$$\textbf{If } x_n \leq y_n \textbf{ for all } n \textbf{ sufficiently large, then } x \leq y. \tag{2.7}$$

Proof: $|x_n - x| < \epsilon$, so $-\epsilon < x_n - x < \epsilon$, implying $-x > -x_n - \epsilon$. Similarly, $|y_n - y| < \epsilon$, or $-\epsilon < y_n - y < \epsilon$, implying $y > y_n - \epsilon$. Adding the two inequalities gives

$$y - x \geq (y_n - \epsilon) - (x_n + \epsilon) = (y_n - x_n) - 2\epsilon \geq -2\epsilon.$$

As ϵ is arbitrary, it follows that $y - x \geq 0$.

Theorem: Let $\{a_n\}$ be a convergent sequence with $\lim a_n = a$, and suppose that $a_n \geq 0$ for all $n \in \mathbb{N}$. Then

$$a \geq 0. \tag{2.8}$$

Proof: This follows directly from (2.7). We can also argue as follows: Suppose to the contrary that $a < 0$, and let $\varepsilon = |a|/2$. The interval $(a - \varepsilon, a + \varepsilon)$ contains no a_n, i.e., if $a_n \in (a - \varepsilon, a + \varepsilon)$ then $a_n < a + \varepsilon < 0$ which is a contradiction. Thus, $a \geq 0$.

Informally, the limit of a function at a particular point, say x, is the value that $f(x)$ approaches, but need not assume at x. For example, $\lim_{x\to 0}(\sin x)/x = 1$, even though the ratio is not defined at $x = 0$. Formally, as instigated in 1821 by Cauchy:

Definition (The δ-ϵ definition of left- and right-hand limits of functions): The function $f : A \subset \mathbb{R} \to \mathbb{R}$ has the *right-hand limit* L *at* c, if, for $c \in \mathbb{R}$, $\forall \epsilon > 0$, $\exists \delta > 0$ such that

$$x \in (c, c+\delta) \cap A \;\Rightarrow\; |f(x) - L| < \epsilon, \tag{2.9}$$

for which we write $L = \lim_{x \to c^+} f(x)$. Likewise, f has the *left-hand limit* L *at* c, if, for $c \in \mathbb{R}$, $\forall \epsilon > 0$, $\exists \delta > 0$ such that

$$x \in (c-\delta, c) \cap A \;\Rightarrow\; |f(x) - L| < \epsilon, \tag{2.10}$$

denoted $L = \lim_{x \to c^-} f(x)$. Observe in both (2.9) and (2.10), point c is not necessarily a member of domain A.

Definition (limit of a function): Using the notation in the previous definition, if $\lim_{x \to c^-} f(x)$ and $\lim_{x \to c^+} f(x)$ exist and coincide, then L is the *limit of* f *at* c, denoted $L = \lim_{x \to c} f(x)$.

We will often use the following equivalent definition:

Definition (Sequential definition of function limit): For function $f : A \subset \mathbb{R} \to \mathbb{R}$ and sequence $\{x_n\} \in A$ such that $x_n \to c$, if sequence $\{f(x_n)\}$ converges as in (2.1), so that $L = \lim_{n \to \infty} f(x_n)$ exists, then L is the *limit of* f *at* c, i.e.,

$$L = \lim_{x \to c} f(x) \iff L = \lim_{n \to \infty} f(x_n). \tag{2.11}$$

Of course, not all limits are finite. We write $\lim_{x \to c^+} f(x) = \infty$ if, $\forall M \in \mathbb{R}$, $\exists\, \delta > 0$ such that $f(x) > M$ for every $x \in (c, c+\delta)$; and $\lim_{x \to c^-} f(x) = \infty$ if, $\forall M \in \mathbb{R}$, $\exists\, \delta > 0$ such that $f(x) > M$ for every $x \in (c-\delta, c)$. Similar definitions hold for $\lim_{x \to c^+} f(x) = -\infty$ and $\lim_{x \to c^-} f(x) = -\infty$. As with a finite limit, if $\lim_{x \to c^+} f(x) = \lim_{x \to c^-} f(x) = \pm\infty$, then we write $\lim_{x \to c} f(x) = \pm\infty$. Lastly, we write $\lim_{x \to \infty} f(x) = L$ if, for each $\epsilon > 0$, $\exists\, x_0$ such that $|f(x) - L| < \epsilon$ for all $x > x_0$, and $\lim_{x \to -\infty} f(x) = L$ if, for each $\epsilon > 0$, $\exists\, x_0$ such that $|f(x) - L| < \epsilon$ for all $x < x_0$. As a shorthand, let $f(\infty) := \lim_{x \to \infty} f(x)$ and $f(-\infty) := \lim_{x \to -\infty} f(x)$. If $f(\infty) = f(-\infty)$, then we take $f(\pm\infty) := f(\infty) = f(-\infty)$.

Theorem: Let f and g be functions whose domain contains the point c and such that $\lim_{x \to c} f(x) = L$ and $\lim_{x \to c} g(x) = M$. Then, for constant values $k_1, k_2 \in \mathbb{R}$,

$$\lim_{x \to c} [k_1 f(x) + k_2 g(x)] = k_1 L + k_2 M, \tag{2.12}$$

$$\lim_{x \to c} f(x) g(x) = LM, \tag{2.13}$$

$$\lim_{x \to c} f(x)/g(x) = L/M, \text{ if } M \neq 0, \tag{2.14}$$

$$\text{if } g(x) \leq f(x), \text{ then } M \leq L. \tag{2.15}$$

The proof of (2.12) is a simple application of the triangle inequality (1.7). For (2.14), see, e.g., Stoll, Thm 3.2.1(c)). The proof of (2.15) follows from (2.7). For (2.13), the core of the proof involves considering sequences, say $\{a_n\}$ and $\{b_n\}$, that converge to a and b respectively, and writing

$$|a_n b_n - ab| = |(a_n b_n - a_n b) + (a_n b - ab)| \leq |a_n|\,|b_n - b| + |b|\,|a_n - a|\,. \tag{2.16}$$

By taking n large enough, both of the terms on the rhs of (2.16) can be made arbitrarily small. Now use the sequential limit definition (2.11).

For the limit of a composition of functions, let $b = \lim_{x\to a} f(x)$ and $L = \lim_{y\to b} g(y)$. Then

$$\lim_{x\to a} g(f(x)) = L. \tag{2.17}$$

Example 2.1 *Compute*

$$\lim_{h\to 0} \frac{(e^h - 1)(\sin t^2)}{h^2}.$$

Using (2.13)*, we can separately compute* $\lim_{h\to 0}(e^h - 1)/h$ *and* $\lim_{t\to 0}(\sin t^2)/t$*. For the former, use of l'Hôpital's rule* (2.44) *yields the limit to be 1. Alternatively, from power series expansion of the exponential function* (2.241) *applied to the numerator,* $\lim_{h\to 0}(e^h - 1)/h = \lim_{h\to 0}(1 + h/2 + \cdots)$*, yielding 1. For the latter, with* $x = t^2 > 0$*,* $\lim_{t\to 0}(\sin t^2)/t = \lim_{x\to 0} \sin(x)/\sqrt{x} = \lim_{x\to 0} \big[\sin(x)/x\big] \times \big[x/\sqrt{x}\big] = \lim_{x\to 0}(\sin(x)/x) \times \lim_{x\to 0} \sqrt{x} = 1\times 0 = 0$*, having used* (2.56)*. The desired limit is thus* 1×0*.* ■

Definition: Let f be a function with domain $A \subset \mathbb{R}$ and $a \in A$. (Note that, without specification of the codomain, it is understood to be $\mathbb{R}$, which is sometimes also stated as saying "let f be a real-valued function".) If $\lim_{x\to a^+} f(x) = f(a)$, then f is said to be *continuous on the right at* a; and if $\lim_{x\to a^-} f(x) = f(a)$, then f is *continuous on the left at* a. We have continuity at point a when both of these conditions hold:

$$f \textbf{ is continuous at } a \textbf{ if } \lim_{x\to a} f(x) = f(a). \tag{2.18}$$

Definition: If f is continuous at each point $a \in S \subset A \subset \mathbb{R}$, then f is *continuous on* S, in which case we also say that f is of *class* $\mathcal{C}^0$ on S, or $f \in \mathcal{C}^0(S)$. Often, subset S will be an interval, say (a, b) or $[a, b]$, in which case we write $f \in \mathcal{C}^0(a, b)$ and $f \in \mathcal{C}^0[a, b]$, respectively. If f is continuous on (its whole domain) A, then we say f is continuous, or that f is of class $\mathcal{C}^0$, or $f \in \mathcal{C}^0$.

From the above definitions, we can express the limit result for continuous functions as follows. If $f : A \subset \mathbb{R} \to \mathbb{R}$ and $f \in \mathcal{C}^0(S)$ for $S \subset A$, then

$$\forall a \in S, \qquad \lim_{x\to a} f(x) = f\left(\lim_{x\to a} x\right) = f(a). \tag{2.19}$$

Theorem: Let function $f : D \to \mathbb{R}$ be continuous. Define $|f| : D \to \mathbb{R}$ by $|f|(x) = |f(x)|$. Then $|f|$ is also continuous.

> Proof: Let $\{x_n\}$ be a sequence in D such that $x_n \to c$. From the continuity of f and the sequential limit definition (2.11), $f(x_n) \to f(c)$. Result (2.6) implies $|f(x_n)| \to |f(c)|$.

Given two functions $f : D \to \mathbb{R}$ and $g : D \to \mathbb{R}$, we define the sum $f + g : D \to \mathbb{R}$ and the product $fg : D \to \mathbb{R}$ by $(f + g)(x) \equiv f(x) + g(x)$ and $(fg)(x) \equiv f(x)g(x)$, $\forall x \in D$. Moreover, if $g(x) \neq 0$ for all x in D, the quotient $f/g : D \to \mathbb{R}$ is defined by

$$(f/g)(x) \equiv \frac{f(x)}{g(x)} \quad \text{for all } x \text{ in } D.$$

The following theorem is an analog, and also a consequence, of the sum, product, and quotient properties of convergent sequences.

Theorem: Suppose that the functions $f : D \to \mathbb{R}$ and $g : D \to \mathbb{R}$ are continuous at the point x_0 in D. Then the sum

$$f + g : D \to \mathbb{R} \textbf{ is continuous at } x_0; \tag{2.20}$$

the product

$$fg : D \to \mathbb{R} \textbf{ is continuous at } x_0; \tag{2.21}$$

and, if $g(x) \neq 0$ for all x in D, the quotient

$$f/g : D \to \mathbb{R} \textbf{ is continuous at } x_0. \tag{2.22}$$

Proof: Let $\{x_n\}$ be a sequence in D that converges to x_0. From definition (2.19), and the sequential limit definition (2.11), $\lim_{n\to\infty} f(x_n) = f(x_0)$ and $\lim_{n\to\infty} g(x_n) = g(x_0)$. Now observe that (2.12) implies

$$\lim_{n\to\infty} [f(x_n) + g(x_n)] = f(x_0) + g(x_0);$$

(2.13) implies

$$\lim_{n\to\infty} [f(x_n)\, g(x_n)] = f(x_0)\, g(x_0);$$

and, if $g(x) \neq 0$ for all x in D, (2.14) implies

$$\lim_{n\to\infty} \frac{f(x_n)}{g(x_n)} = \frac{f(x_0)}{g(x_0)}.$$

Results (2.20), (2.21), and (2.22) follow.

Theorem: Let $f : D \to \mathbb{R}$ be a continuous function on (nonempty) interval $D = (a, b) \in \mathbb{R}$. Let $c \in (a, b)$ such that $f(c) > 0$. Prove:

$$\exists \delta > 0 \textbf{ such that } f(x) > 0 \textbf{ for } x \in (c - \delta, c + \delta). \tag{2.23}$$

Likewise, if $f(c) < 0$, then there is $\delta > 0$ such that $f(x) < 0$ whenever $x \in D$ and $|x - c| < \delta$.

Proof: Suppose to the contrary that no such δ exists. Then, for every $\delta > 0$, there exists $x \in (c-\delta, c+\delta)$ such that $f(x) \leq 0$. If we take $\delta = 1/n$, we obtain a sequence x_n in $(c - 1/n, c + 1/n)$ with $f(x_n) \leq 0$. The inequality $|x_n - c| < 1/n$ shows that the sequence x_n converges to c, and the continuity of f implies that the sequence $f(x_n)$ converges to $f(c)$. From the previous question and that $f(x_n) \leq 0$, $f(c) \leq 0$. This contradicts the assumption that $f(c) > 0$.

We wish to devise an example to show that the converse of the previous theorem is not true. Hint: This means you need to demonstrate a continuous function f such that

$$\forall x \in \{x : 0 < |x - c| < \delta\}, \quad f(x) > 0, \quad f(c) \leq 0.$$

An example is: Take $f(x) = x^2$, $c = 0$.

Theorem: Let $f : D \to \mathbb{R}$ be a continuous function at a point c in $(a, b) \subset D$. Prove:

$$\textbf{If } f(x) \geq 0 \textbf{ on } (a, c) \cup (c, b) \textbf{ then } f(c) \geq 0. \tag{2.24}$$

Proof: Let $\{a_n\}$ be a sequence in (a,b) converging to c, and $a_n \neq c$. Then $f(a_n) \geq 0$, and the result (2.8) implies that $\lim f(a_n) \geq 0$. Since f is continuous at c, this means that $f(c) \geq 0$.

An important result is the continuity of composite functions: Let $f : A \to B$ and $g : B \to C$ be continuous. Then $g \circ f : A \to C$ is continuous. More precisely, if f is continuous at $a \in A$, and g is continuous at $b = f(a) \in B$, then

$$\lim_{x \to a} g(f(x)) = g\left(\lim_{x \to a} f(x)\right). \tag{2.25}$$

We defined continuity of function $f : A \subset \mathbb{R} \to \mathbb{R}$ at point $a \in A$ in (2.18) as being when $\lim_{x \to a} f(x) = f(a)$. We now give another popular definition, with their equivalence being shown below.

Definition: Let f be a function with domain $A \subset \mathbb{R}$. Function f is continuous at $a \in A$ if, given $\epsilon > 0$, $\exists \delta > 0$ such that, if $|x - a| < \delta$ and $x \in A$, then $|f(x) - f(a)| < \epsilon$.

Its value is seen when contrasting it with a definition of *uniform continuity*:

Definition: Let f be a function with domain A and let $[a,b] \subset A$ be a closed, finite interval. Function f is uniformly continuous on $[a,b]$ if the condition holds: For a given $\epsilon > 0$, $\exists \delta > 0$ such that, if $x, y \in [a,b]$, and $|x - y| < \delta$, then $|f(x) - f(y)| < \epsilon$.

Note crucially that, with uniform continuity, δ does not depend on the choice of $x \in [a,b]$.

The reader might guess that there is a comparable, equivalent definition of uniform continuity in terms of limits of sequences. This is true, and important, and in fact we take this as our formal definition:

Definition: Let $D \subseteq \mathbb{R}$ and let $f : D \to \mathbb{R}$ be a function. We say that f is uniformly continuous on D if $(x_n), (y_n)$ any sequences in D and $x_n - y_n \to 0 \Longrightarrow f(x_n) - f(y_n) \to 0$.

The equivalences of the two forms of definitions, for both continuity, and uniform continuity, are proved in most all beginning books on real analysis. The proofs are instructive, and we give them here, as presented in (the magnificent) Ghorpade and Limaye (2018, Propositions 3.8 and 3.22), along with some examples and further results.

Theorem: Let $D \subseteq \mathbb{R}$, $c \in D$, and let $f : D \to \mathbb{R}$ be a function. Then f is continuous at c if and only if f satisfies the following $\epsilon - \delta$ condition: $\forall \epsilon > 0$, $\exists \delta > 0$ such that

$$x \in D \text{ and } |x - c| < \delta \Longrightarrow |f(x) - f(c)| < \epsilon. \tag{2.26}$$

Proof: Let f be continuous at c. Suppose the $\epsilon - \delta$ condition does not hold. This means that there is $\epsilon > 0$ such that for every $\delta > 0$, there is $x \in D$ satisfying

$$|x - c| < \delta, \quad \text{but} \quad |f(x) - f(c)| \geq \epsilon.$$

Then there is a sequence (x_n) in D such that $|x_n - c| < 1/n$, but $|\, f(x_n) - f(c) \,| \geq \epsilon$ for all $n \in \mathbb{N}$. But then $x_n \to c$ and $f(x_n) \nrightarrow f(c)$. This contradicts the continuity of f at c.

Conversely, assume the $\epsilon - \delta$ condition. Let (x_n) be any sequence in D such that $x_n \to c$. Let $\epsilon > 0$ be given. Then there is $\delta > 0$ such that

$$x \in D \text{ and } |x - c| < \delta \Longrightarrow |f(x) - f(c)| < \epsilon.$$

Since $x_n \to c$, there is $n_0 \in \mathbb{N}$ such that $|x_n - c| < \delta$ for all $n \geq n_0$. Hence $|f(x_n) - f(c)| < \epsilon$ for all $n \geq n_0$. Thus $f(x_n) \to f(c)$. This shows that f is continuous at c.

Theorem: Let $D \subseteq \mathbb{R}$ and let $f : D \to \mathbb{R}$ be a function. Then f is uniformly continuous on D if and only if f satisfies the following uniform $\epsilon - \delta$ condition: For every $\epsilon > 0$, there is $\delta > 0$ such that

$$x, y \in D \text{ and } |x - y| < \delta \Longrightarrow |f(x) - f(y)| < \epsilon.$$

Proof: Let f be uniformly continuous on D. Suppose there is $\epsilon > 0$ such that for every $\delta > 0$, there are x and y in D such that $|x - y| < \delta$, but $|f(x) - f(y)| \geq \epsilon$. Considering $\delta := 1/n$ for $n \in \mathbb{N}$, we obtain sequences (x_n) and (y_n) in D such that $|x_n - y_n| < 1/n$ but $|f(x_n) - f(y_n)| \geq \epsilon$ for all $n \in \mathbb{N}$. Then $x_n - y_n \to 0$, but $f(x_n) - f(y_n) \nrightarrow 0$. This contradicts the assumption that f is uniformly continuous on D.

Conversely, assume that the uniform $\epsilon - \delta$ condition holds. Let (x_n) and (y_n) be any sequences in D such that $x_n - y_n \to 0$. Let $\epsilon > 0$ be given. Then there is $\delta > 0$ such that $|f(x) - f(y)| < \epsilon$, whenever $x, y \in D$ and $|x - y| < \delta$. Since $x_n - y_n \to 0$, we can find $n_0 \in \mathbb{N}$ such that $|x_n - y_n| < \delta$ for all $n \geq n_0$. But then $|f(x_n) - f(y_n)| < \epsilon$ for all $n \geq n_0$. Thus $f(x_n) - f(y_n) \to 0$. Hence f is uniformly continuous on D.

Theorem: Show that uniformly continuous functions defined on the same domain form a vector space. That means, if $f, g : D \to \mathbb{R}$ are uniformly continuous functions, then $cf + dg$ is uniformly continuous, where $c, d \in \mathbb{R}$.

Proof: We show additivity and scalar multiplication (homogeneity) separately:

Additivity: Given $\epsilon > 0$ there are $\delta_1 > 0$ and $\delta_2 > 0$ such that for any $x, y \in D$, if $|x - y| < \delta_1$, then $|f(x) - f(y)| < \epsilon/2$ and if $|x - y| < \delta_2$, then $|g(x) - g(y)| < \epsilon/2$. Therefore, if $|x - y| < \min\{\delta_1, \delta_2\}$, then, from the triangle inequality,

$$|(f + g)(x) - (f + g)(y)| \leq |f(x) - f(y)| + |g(x) - g(y)| < \epsilon/2 + \epsilon/2 = \epsilon.$$

Homogeneity: By uniform continuity of f, we have that, given $\epsilon > 0$, there is a $\delta > 0$ such that, for any $x, y \in D$, if $|x - y| < \delta$, then $|f(x) - f(y)| < \epsilon/|c|$, for $c \in \mathbb{R} \setminus \{0\}$. Therefore, if $|x - y| < \delta$,

$$|cf(x) - cf(y)| \leq |c||f(x) - f(y)| < |c|\epsilon/|c| = \epsilon.$$

Combining the two statements gives the result.

The next result (Ghorpade and Limaye, Prop. 3.20) is very important in analysis. We will use it in proving (5.24). Its proof invokes the Bolzano–Weierstrass Theorem, and can be skipped by readers not familiar with this material. Recall that, if f is a continuous function on its domain D, we write $f \in \mathcal{C}^0(D)$.

Theorem: Let $D \subseteq \mathbb{R}$. Every uniformly continuous function on D is continuous on D. Moreover, if D is a closed and bounded set, and $f \in \mathcal{C}^0(D)$, then

$$f \textbf{ is uniformly continuous on } D. \tag{2.27}$$

Proof: Let $f : D \to \mathbb{R}$ be given. First assume that f is uniformly continuous on D. If $c \in D$ and (x_n) is any sequence in D such that $x_n \to c$, then let $y_n := c$ for all $n \in \mathbb{N}$. Since $x_n - y_n \to 0$, we obtain $f(x_n) - f(c) = f(x_n) - f(y_n) \to 0$, that is, $f(x_n) \to f(c)$. Thus f is continuous at c. Since this holds for every $c \in D$, f is continuous on D.

Now assume that D is a closed and bounded set and f is continuous on D. Suppose f is not uniformly continuous on D. Then there are sequences (x_n) and (y_n) in D such that $x_n - y_n \to 0$, but $f(x_n) - f(y_n) \nrightarrow 0$. Consequently, there exist $\epsilon > 0$ and positive integers $n_1 < n_2 < \cdots$ such that $|f(x_{n_k}) - f(y_{n_k})| \geq \epsilon$ for all $k \in \mathbb{N}$. Since D is a bounded set, the sequence $\{x_{n_k}\}$ is bounded. By the Bolzano-Weierstrass Theorem, it has a convergent subsequence, say $\{x_{n_{k_j}}\}$. Let us denote the sequences $\{x_{n_{k_j}}\}$ and $\{y_{n_{k_j}}\}$ by $(\tilde{x}_j)$ and $(\tilde{y}_j)$ for simplicity. Let $\tilde{x}_j \to c$. Then $c \in D$, since D is a closed set. Because $x_n - y_n \to 0$, we see that $\tilde{x}_j - \tilde{y}_j \to 0$ and hence $\tilde{y}_j \to c$ as well. Since f is continuous at c, we obtain $f(\tilde{x}_j) \to f(c)$ and $f(\tilde{y}_j) \to f(c)$. Thus

$$f(\tilde{x}_j) - f(\tilde{y}_j) \to f(c) - f(c) = 0.$$

But this is a contradiction, since $|f(\tilde{x}_j) - f(\tilde{y}_j)| \geq \epsilon$ for all $j \in \mathbb{N}$. Hence f is uniformly continuous on D.

The notions of uniform continuity and uniform convergence (the latter discussed in §2.5 below) play a major role in analysis. One example, of many, and which we will use, is in the construction of the Riemann integral in §2.4.1.

We now gather some fundamental results, stated without proof; the reason being, their proofs are most easily conducted invoking compactness, which we do not cover in this class. Let $I = [a, b]$ be a closed, bounded interval. (This in fact means that I is compact.) Let f be a continuous function on I, i.e., $f \in \mathcal{C}^0[a, b]$. Then:

- The image of $f \in \mathcal{C}^0(I)$, $I = [a, b]$, forms a closed, bounded subset of $\mathbb{R}$, i.e.,

$$\forall x \in I, \exists m, M \in \mathbb{R} \textbf{ such that } m \leq f(x) \leq M. \tag{2.28}$$

- Function f assumes minimum and maximum values on I, i.e.,[5]

$$f \in \mathcal{C}^0(I), I = [a, b] \implies \exists x_0, x_1 \in I \textbf{ s.t. } \forall x \in I, f(x_0) \leq f(x) \leq f(x_1). \tag{2.29}$$

- (*Intermediate Value Theorem*) Let $f : D \subset \mathbb{R} \to \mathbb{R}$, $I = [a, b] \subset D$, $f \in \mathcal{C}^0(I)$, and $\alpha = f(a)$ and $\beta = f(b)$.

$$\forall \gamma : \alpha < \gamma < \beta, \ \exists c \in (a, b) \textbf{ such that } f(c) = \gamma. \tag{2.30}$$

- As stated in (2.27) and proved there,

$$f \in \mathcal{C}^0(I), I = [a, b] \implies f \textbf{ is uniformly continuous on } I. \tag{2.31}$$

These four facts together constitute what Pugh (2002, p. 39) argues could rightfully be called the *Fundamental Theorem of Continuous Functions*.

[5] According to Petrovic, Advanced Calculus: Theory and Practice, 2nd ed., 2020, p. 99, this is known as "The Maximum Theorem", "The Extreme Value Theorem", and Weierstrass called it "The Principal Theorem" in his lectures in 1861. The result was originally proved by Bolzano, but his proof was not published until 1930. The first publication was by Cantor in 1870.

2.2 Differentiation

2.2.1 Definitions and Techniques

Let $f \in \mathcal{C}^0(I)$, where I is an interval of nonzero length. If the *Newton quotient*

$$\lim_{h\to 0} \frac{f(x+h) - f(x)}{h} \tag{2.32}$$

exists for $x \in I$, then f is *differentiable* at x, the limit is the *derivative* of f at x, and is denoted $f'(x)$ or df/dx. Similar to the notation for continuity, if f is differentiable at each point in I, then f is differentiable on I, and if f is differentiable on its domain, then f is differentiable. If f is differentiable and $f'(x)$ is a continuous function of x, then f is *continuously differentiable*, and is of *class* $\mathcal{C}^1$.

Observe that, for h small,

$$f'(x) \approx \frac{f(x+h) - f(x)}{h} \quad \text{or} \quad f(x+h) \approx f(x) + hf'(x), \tag{2.33}$$

which, for constant x, is a linear function in h. By letting $h = y - x$, (2.32) can be equivalently written as

$$\lim_{y\to x} \frac{f(y) - f(x)}{y - x}, \tag{2.34}$$

which is sometimes more convenient to work with.

Lemma (Fundamental lemma of differentiation) This lemma makes the notion more precise that a differentiable function can be approximated at each point in (the interior of) its domain by a linear function whose slope is the derivative at that point. As in Protter and Morrey (1991, p. 85), let f be differentiable at the point x. Then there exists a function η defined on an interval about zero such that

$$f(x+h) - f(x) = \big[f'(x) + \eta(h)\big] \cdot h, \tag{2.35}$$

and η is continuous at zero, with $\eta(0) = 0$. The proof follows by solving (2.35) for $\eta(h)$ and defining $\eta(0) = 0$, i.e.,

$$\eta(h) := \frac{1}{h}\big[f(x+h) - f(x)\big] - f'(x), \quad h \neq 0, \qquad \eta(0) := 0.$$

As f is differentiable at x, $\lim_{h\to 0} \eta(h) = 0$, so that η is continuous at zero.

Example 2.2 *From the definition of limit,* $\lim_{h\to 0} 0/h = 0$*, so that the derivative of* $f(x) = k$ *for some constant* k *is zero. For* $f(x) = x$*, it is easy to see from (2.32) that* $f'(x) = 1$*. For* $f(x) = x^2$*,*

$$\lim_{h\to 0} \frac{(x+h)^2 - x^2}{h} = \lim_{h\to 0} (2x + h) = 2x.$$

Now consider $f(x) = x^n$ *for* $n \in \mathbb{N}$*. The binomial theorem implies*

$$f(x+h) = (x+h)^n = \sum_{i=0}^{n} \binom{n}{i} x^{n-i} h^i = x^n + nhx^{n-1} + \cdots + h^n,$$

so that

$$\lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} \left(nx^{n-1} + \cdots + h^{n-1}\right) = nx^{n-1}. \tag{2.36}$$

Now let $f(x) = x^{-n}$ *for* $n \in \mathbb{N}$. *It is easy to verify that, for any* $x \neq 0$ *and* $y \neq 0$,

$$f(y) - f(x) = y^{-n} - x^{-n} = \frac{x^n - y^n}{x^n y^n} = (y - x)\left[-\frac{y^{n-1} + y^{n-2}x + \cdots + x^{n-1}}{x^n y^n}\right],$$

so that (2.34) implies

$$\lim_{y \to x} \frac{f(y) - f(x)}{y - x} = -\lim_{y \to x}\left[\frac{y^{n-1} + y^{n-2}x + \cdots + x^{n-1}}{x^n y^n}\right] = -\frac{nx^{n-1}}{x^{2n}} = -nx^{-n-1}.$$

Thus, for $f(x) = x^n$ *with* $n \in \mathbb{Z}$, $f'(x) = nx^{n-1}$. ■

Assume for functions f and g defined on I that $f'(x)$ and $g'(x)$ exist on I. Then

$$\textbf{(sum rule)} \qquad (f+g)'(x) = f'(x) + g'(x), \tag{2.37}$$

$$\textbf{(product rule)} \qquad (fg)'(x) = f(x)\,g'(x) + g(x)\,f'(x), \tag{2.38}$$

$$\textbf{(quotient rule)} \qquad (f/g)'(x) = \frac{g(x)\,f'(x) - f(x)\,g'(x)}{[g(x)]^2}, \quad g(x) \neq 0, \tag{2.39}$$

$$\textbf{(chain rule)} \qquad (g \circ f)'(x) = g'(f(x))\,f'(x). \tag{2.40}$$

The chain rule is proven in all beginning real analysis books. It is simple to prove using the fundamental lemma of differentiation (2.35); see, e.g., Protter and Morrey (1991, p. 85), or Ghorpade and Limaye (2018, Proposition 4.10).

Remark 1: Simply, but usefully, from (2.37); and (2.38) with $f(x) = -1$,

$$(f - g)'(x) = f'(x) - g'(x). \tag{2.41}$$

Remark 2: The set of differential functions forms a vector space. That means, if f and g are differentiable functions on their domain D, and $a, b \in \mathbb{R}$, then functions $(af) : D \to \mathbb{R}$, with $(af)(x) := af(x)$, and $(f+g) : D \to \mathbb{R}$, with $(f+g)(x) := f(x)+g(x)$, are differentiable; these properties being called homogeneity and linearity, respectively.

Remark 3: With $y = f(x)$ and $z = g(y)$, the usual mnemonic for the chain rule is

$$\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}.$$

Example 2.3 *Result (2.36) for* $n \in \mathbb{N}$ *could also be established by using an induction argument: Let* $f(x) = x^n$ *and assume* $f'(x) = nx^{n-1}$. *It holds for* $n = 1$; *induction and the product rule imply for* $f(x) = x^{n+1} = x^n \cdot x$ *that* $f'(x) = x^n \cdot 1 + x \cdot nx^{n-1} = (n+1)\,x^n$. ■

Theorem:

If f is differentiable at a, then f is continuous at a. (2.42)

Proof: This is seen by taking limits of

$$f(x) = \frac{f(x) - f(a)}{x - a}(x - a) + f(a),$$

which, using (2.12) and (2.13), gives

$$\lim_{x \to a} f(x) = f'(a) \cdot 0 + f(a) = f(a),$$

and recalling the definition of continuity, e.g., (2.19).

The function $f(x) = |x|$ at $x = 0$ is the showcase example that continuity does not imply differentiability.

Proposition: Differentiability of f does not imply that f' is continuous.

> Proof: We just need a counterexample. A popular one takes $f(x) = x^2 \sin(1/x)$ for $x \neq 0$ and $f(0) = 0$. Then
>
> $$f'(x) = 2x \sin\left(\frac{1}{x}\right) - \cos\left(\frac{1}{x}\right), \quad x \neq 0,$$
>
> and $\lim_{x\to 0} f'(x)$ does not exist. But, from the Newton quotient at $x = 0$,
>
> $$\lim_{h\to 0} \frac{f(0+h) - f(0)}{h} = \lim_{h\to 0} h \sin(1/h) = 0,$$
>
> so that $f'(0) = 0$. Thus, $f'(x)$ is not continuous.

What is true is that *uniform differentiability* implies uniform continuity of the derivative, as discussed next. I took this from Estep (2002, §32.4), and it is also stated as an exercise in Stoll (2021, p. 205, # 16).

Definition: A function f is said to be uniformly differentiable on an interval $[a, b]$ if, $\forall \epsilon > 0$, $\exists \delta > 0$ such that

$$\left| \frac{f(y) - f(x)}{y - x} - f'(x) \right| < \epsilon, \quad \forall x, y \in [a, b] \quad \text{with} \quad |x - y| < \delta.$$

Theorem: If f is uniformly differentiable on $[a, b]$, then $f'(x)$ is uniformly continuous on $[a, b]$.

> Proof: If f is uniformly differentiable, then for $x, y \in [a, b]$ and $\epsilon > 0$, we can find a $\delta > 0$ such that, for $|x - y| < \delta$,
>
> $$\begin{aligned} |f'(y) - f'(x)| &= \left| f'(y) - \frac{f(y) - f(x)}{y - x} + \frac{f(y) - f(x)}{y - x} - f'(x) \right| \\ &\leq \left| f'(y) - \frac{f(y) - f(x)}{y - x} \right| + \left| \frac{f(y) - f(x)}{y - x} - f'(x) \right| < 2\epsilon. \end{aligned}$$
>
> Thus, $f'(x)$ is uniformly continuous on $[a, b]$.

Example 2.4 *Function $f(x) = x^2$ is uniformly differentiable on any bounded interval $[a, b]$. The function $f(x) = 1/x$ is differentiable on $(0, 1)$, but is not uniformly differentiable on $(0, 1)$.* ■

Theorem (Fermat): Let f be defined on an interval $[a, b]$ and suppose that it attains its greatest or its smallest value at a point $c \in (a, b)$.

$$\textbf{If } f \textbf{ is differentiable at } c\textbf{, then } f'(c) = 0. \tag{2.43}$$

Proof: We will assume that f attains its greatest value at $c \in (a,b)$, i.e., that $f(x) \leq f(c)$ for all $x \in [a,b]$. If $x < c$ then

$$\frac{f(x)-f(c)}{x-c} \geq 0$$

and (2.24) implies that $f'(c) \geq 0$. On the other hand, if $x > c$ then

$$\frac{f(x)-f(c)}{x-c} \leq 0$$

so $f'(c) \leq 0$. Combining these two inequalities, $f'(c) \geq 0$ and $f'(c) \leq 0$, we obtain that $f'(c) = 0$.

Of great use is *l'Hôpital's rule*[6] for evaluating *indeterminate forms or ratios*:

Theorem (l'Hôpital's rule, 0/0 case): Let f and g, and their first derivatives, be continuous functions on (a,b). If $\lim_{x\to a^+} f(x) = \lim_{x\to a^+} g(x) = 0$ and $\lim_{x\to a^+} f'(x)/g'(x) = L$, then

$$\lim_{x\to a^+} f(x)/g(x) = L. \tag{2.44}$$

Most students remember this very handy result, but few can intuitively justify it. Most real analysis textbooks give the rigorous proof, and also discuss and prove the ∞/∞ case. We give a "heuristic justification" that is easy to remember.

Assume f and g are continuous at a, so that $f(a) = g(a) = 0$. Using (2.33) gives

$$\lim_{x\to a^+} \frac{f(x)}{g(x)} = \lim_{h\to 0} \frac{f(a+h)}{g(a+h)} \approx \lim_{h\to 0} \frac{f(a)+hf'(a)}{g(a)+hg'(a)} = \frac{f'(a)}{g'(a)} = \lim_{x\to a^+} \frac{f'(x)}{g'(x)}.$$

A different "rough proof" of l'Hôpital's rule is given by Pugh (2002, p. 143) and Stoll, 2021, p. 212, #2.

A similar result, also referred to as l'Hôpital's rule, holds for $x \to b^-$, and for $x \to \infty$; and also for the case when $\lim_{x\to a^+} f(x) = \lim_{x\to a^+} g(x) = \infty$.

Example 2.5 *(Petrovic, Advanced Calculus: Theory and Practice, 2nd ed., 2020, Example 4.6.4) Determine*

$$\lim_{x\to 1} x^{\frac{1}{1-x}}.$$

This limit is of the form (1^∞)*, so we cannot apply l'Hôpital's rule directly. Since* $x = e^{\ln x}$ *when* $x > 0$*, and since* $x \to 1$ *means that we can assume that* $x > 0$*, we have*

$$x^{\frac{1}{1-x}} = \exp\left(\ln x^{\frac{1}{1-x}}\right) = \exp\left(\frac{\ln x}{1-x}\right).$$

When $x \to 1$*, the exponent* $\ln x/(1-x)$ *is of the form* $\left(\frac{0}{0}\right)$*, so we can apply l'Hôpital's rule. Further,* $(1-x)' = -1 \neq 0$ *so*

$$\lim_{x\to 1} \frac{\ln x}{1-x} = \lim_{x\to 1} \frac{1/x}{-1} = -1,$$

[6]Named after Guillaume François Antoine Marquis de l'Hôpital (1661–1704), who was taught calculus by Johann Bernoulli (for a high price), and wrote the first calculus textbook (1696) based on Bernoulli's notes, in which the result appeared. Not surprisingly, l'Hôpital's rule was also known to Bernoulli (confirmed in Basel, 1922, with the discovery of certain written documents).

and we obtain that $\lim_{x\to 1} x^{\frac{1}{1-x}} = e^{-1}$. ■

Let f be a strictly increasing continuous function on a closed interval I. Then, from the intermediate value theorem (2.30), the image of f is also a closed interval. Function f on I is said to be invertible, or is a bijection, meaning it is both injective and surjective (see page 4). The *inverse function* $g = f^{-1}$ is defined as the function such that $(g \circ f)(x) = x$ and $(f \circ g)(y) = y$. It is also continuous and strictly increasing. If f is also differentiable in the interior of I with $f'(x) > 0$, then a fundamental result is that

$$g'(y) = \frac{1}{f'(x)} = \frac{1}{f'(g(y))}. \tag{2.45}$$

We prove a simpler version of this. It assumes existence of the derivative of f^{-1}.

Lemma: Let $X, Y \subset \mathbb{R}$, and let $f : X \to Y$ be an invertible function, with inverse $f^{-1} : Y \to X$. Suppose that $x_0 \in X$ and $y_0 \in Y$ are limit points of X, Y, respectively, such that $y_0 = f(x_0)$. This implies $x_0 = f^{-1}(y_0)$. If f is differentiable at x_0, and f^{-1} is differentiable at y_0, then

$$\left(f^{-1}\right)'(y_0) = \frac{1}{f'(x_0)}.$$

> Proof: (Tao, Analysis I, 4th ed., 2022, p. 226) First note that, if f is the identity function, i.e., $f(x) = x$ for all $x \in X$, then f is differentiable at x_0 and $f'(x_0) = 1$. From the chain rule (2.40),
>
> $$\left(f^{-1} \circ f\right)'(x_0) = \left(f^{-1}\right)'(y_0) f'(x_0).$$
>
> But $f^{-1} \circ f$ is the identity function on X, and hence $(f^{-1} \circ f)'(x_0) = 1$.

(Tao then gives the more general proof, which relaxes the requirement on f^{-1} from differentiability to continuity.)

A useful application involving the arcsin and arctan functions is given below in Example 2.7.

We close this section with two definitions that are of occasional use. We will use them below in proving (2.117).

Definition: (As in Stoll, Def 5.1.2) Let $I \subset \mathbb{R}$ be an interval and let f be a real-valued function with domain I. If $p \in I$ is such that $I \cap (p, \infty) \neq \emptyset$, then the right derivative of f at p, denoted $f'_+(p)$, is defined as

$$f'_+(p) = \lim_{h\to 0^+} \frac{f(p+h) - f(p)}{h}, \tag{2.46}$$

provided the limit exists. Similarly, if $p \in I$ satisfies $(-\infty, p) \cap I \neq \emptyset$, then the left derivative of f at p, denoted $f'_-(p)$, is given by

$$f'_-(p) = \lim_{h\to 0^-} \frac{f(p+h) - f(p)}{h}, \tag{2.47}$$

provided the limit exists.

NOTE: if $I = [a, b]$, the right derivative applies to $p \in [a, b)$, but not for $p = b$, because $I \cap (b, \infty) = \emptyset$. Similar for left derivative.

2.2.2 Trigonometric Functions

We turn now to some fundamental results on limits and derivatives for trigonometric functions. Recall the unit circle geometric representation of sine and cosine, and, from Pythagoras, the fundamental relation $\cos^2(x) + \sin^2(x) = 1$. Of great use are the following relations:

Theorem (Angle sum and difference identities): For $x, y \in \mathbb{R}$,

$$\sin(x+y) = \sin x \cos y + \cos x \sin y, \tag{2.48a}$$
$$\cos(x+y) = \cos x \cos y - \sin x \sin y. \tag{2.48b}$$

These can be demonstrated from a clever graphic, such as in Stillwell, Numbers and Geometry, 1998, §5.3, though I prefer the nice derivation from Kuttler, Calculus of One and Many Variables, p. 59), included here.

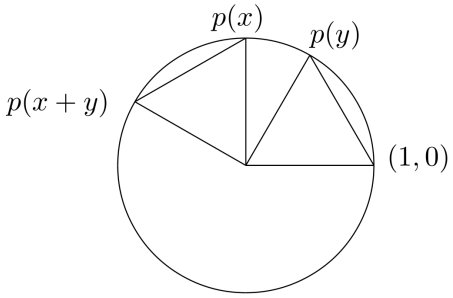


Figure 1: From Kuttler. Unit circle with two inscribed, equal triangles

Theorem: Let $x, y \in \mathbb{R}$. Then

$$\cos(x+y)\cos(x) + \sin(x+y)\sin(x) = \cos(y). \tag{2.49}$$

Proof: Recall that, for a real number z, there is a unique point $p(z)$ on the unit circle and the coordinates of this point are $\cos z$ and $\sin z$. Now it seems geometrically clear from Figure 1 that the length of the arc between $p(x+y)$ and $p(x)$ has the same length as the arc between $p(y)$ and $p(0)$.

Also from geometric reasoning the distance between the points $p(x+y)$ and $p(x)$ must be the same as the distance from $p(y)$ to $p(0)$. In fact, the two triangles have the same angles and the same sides. Writing this in terms of the definition of the trig functions and the distance formula,

$$(\cos(x+y) - \cos x)^2 + (\sin(x+y) - \sin x)^2 = (\cos y - 1)^2 + (\sin y - 0)^2.$$

Expanding, we get

$$\cos^2(x+y) + \cos^2 x - 2\cos(x+y)\cos x + \sin^2(x+y) + \sin^2 x - 2\sin(x+y)\sin x = \cos^2 y - 2\cos y + 1 + \sin^2 y.$$

Now using that $\cos^2 + \sin^2 = 1$,

$$2 - 2\cos(x+y)\cos(x) - 2\sin(x+y)\sin(x) = 2 - 2\cos(y),$$

which gives (2.49).

Continuing from Kuttler's presentation, we now prove (2.48).

Proof: The length of the unit circle is defined as 2π. Thus, for example, $\sin\left(\frac{\pi}{2}\right) = 1$, $\cos\left(\frac{\pi}{2}\right) = 0$. Letting $x = \pi/2$, (2.49) implies (seen also from the unit circle)

$$\sin(y + \pi/2) = \cos y. \tag{2.50}$$

Now let $u = x + y$ and $v = x$. Then (2.49) implies $\cos u \cos v + \sin u \sin v = \cos(u - v)$. Also, from this and the basic relations

$$\cos(-x) = \cos(x) \qquad \text{and} \qquad \sin(-x) = -\sin(x), \tag{2.51}$$

we obtain

$$\begin{aligned}\cos(u + v) &= \cos(u - (-v)) = \cos u \cos(-v) + \sin u \sin(-v)\\ &= \cos u \cos v - \sin u \sin v.\end{aligned}$$

Thus, letting $v = \pi/2$ (and also graphically clear from the unit circle),

$$\cos(u + \pi/2) = -\sin u. \tag{2.52}$$

Then, from (2.50) and (2.52),

$$\begin{aligned}\sin(x + y) &= -\cos\left(x + \frac{\pi}{2} + y\right)\\ &= -\left[\cos\left(x + \frac{\pi}{2}\right)\cos y - \sin\left(x + \frac{\pi}{2}\right)\sin y\right]\\ &= \sin x \cos y + \sin y \cos x, \qquad \text{and}\\ \sin(x - y) &= \sin x \cos y - \cos x \sin y.\end{aligned}$$

Using (2.48b), $\cos(2x) = \cos(x + x) = \cos(x)\cos(x) - \sin(x)\sin(x) = \cos^2(x) - \sin^2(x)$. From this, we easily obtain two of the useful double-angle formulae,

$$\cos 2x = \cos^2 x - \sin^2 x = (1 - \sin^2 x) - \sin^2 x = 1 - 2\sin^2 x; \tag{2.53}$$

and

$$\cos 2x = \cos^2 x - \sin^2 x = \cos^2 x - (1 - \cos^2 x) = -1 + 2\cos^2 x. \tag{2.54}$$

Let $f(x) = \sin(x)$. Using (2.48a), the derivative of f is

$$\begin{aligned}\frac{d\sin(x)}{dx} &= \lim_{h\to 0}\frac{\sin(x+h) - \sin(x)}{h} = \lim_{h\to 0}\frac{\sin x\ \cos h - \sin x}{h} + \lim_{h\to 0}\frac{\cos x\ \sin h}{h}\\ &= \sin(x)\lim_{h\to 0}\frac{\cos(h) - 1}{h} + \cos(x)\lim_{h\to 0}\frac{\sin(h)}{h}\\ &= \cos(x),\end{aligned} \tag{2.55}$$

where

$$L_s := \lim_{h\to 0}\frac{\sin(h)}{h} = 1 \quad \text{and} \quad \lim_{h\to 0}\frac{\cos(h) - 1}{h} = 0. \tag{2.56}$$

Both limits in (2.56) need to be justified. If we *assume* that L_s is not infinite, then the second

limit in (2.56) is easy to prove: Write

$$\begin{aligned}\frac{\cos(h)-1}{h} &= \frac{h(\cos(h)+1)}{h(\cos(h)+1)}\frac{\cos(h)-1}{h} = \frac{h(\cos^2 h-1)}{h^2(\cos(h)+1)}\\ &= -\left(\frac{\sin h}{h}\right)^2\frac{h}{\cos(h)+1}\end{aligned}$$

using $\cos^2(x)+\sin^2(x)=1$, so that, from (2.13) and because we assumed that $L_s\in\mathbb{R}$,

$$\lim_{h\to 0}\frac{\cos(h)-1}{h} = -\lim_{h\to 0}\left(\frac{\sin h}{h}\right)^2\lim_{h\to 0}\frac{h}{\cos(h)+1} = 0.$$

Using (2.55), along with (2.50) and (2.52), i.e., $\cos(x)=\sin(x+\pi/2)$ and $\sin(x)=-\cos(x+\pi/2)$, the chain rule gives

$$\frac{d\cos x}{dx} = \frac{d\sin(x+\pi/2)}{dx} = \cos(x+\pi/2) = -\sin(x).$$

Students remember that derivative of sine and cosine involve, respectively, cosine and sine, but some forget the signs. To recall them, just think of the unit circle at angle $\theta=0$ and the geometric definition of sine and cosine. A slight increase in θ increases the vertical coordinate (sine) and decreases the horizontal one (cosine).

The easiest way of proving the former limit in (2.56) is using (2.55); it follows trivially by using the derivative of $\sin x$, i.e.,

$$\lim_{h\to 0}\frac{\sin(h)}{h} = \lim_{h\to 0}\frac{\sin h-\sin 0}{h} = \left.\frac{d\sin x}{dx}\right|_{x=0} = \cos 0 = 1.$$

The limits in (2.56) also follow by applying l'Hôpital's rule: For the latter,

$$\lim_{h\to 0}\frac{\cos(h)-1}{h} = \frac{-\sin(h)}{1} = -\sin(0) = 0.$$

The circular logic between (2.55) and (2.56) is obviously not acceptable.[7] The properties of the sine and cosine functions can be correctly, elegantly and easily derived from an algebraic point of view by starting with functions s and c such that

$$s'=c,\ c'=-s,\ s(0)=0\ \text{ and }\ c(0)=1 \tag{2.57}$$

(see e.g., Lang, 1997, §4.3). As definitions one takes

$$\cos(z)=\sum_{k=0}^{\infty}(-1)^k\frac{z^{2k}}{(2k)!}\qquad\text{and}\qquad \sin(z)=\sum_{k=0}^{\infty}(-1)^k\frac{z^{2k+1}}{(2k+1)!}, \tag{2.58}$$

which converge for all $z\in\mathbb{R}$; see §2.5.5 and §2.5.6 below for details. From (2.58), the properties of the trigonometric functions can be inferred, such as $\cos^2(x)+\sin^2(x)=1$, (2.48), (2.50), (2.51), and (2.52). See e.g., Browder (1996, §3.6) or Hijab (1997, §3.5) for details.

[7] Of course, from a geometric point of view, it is essentially obvious that $\lim_{h\to 0}h^{-1}\sin(h)=1$. Let θ be the angle in the first quadrant of the unit circle, measured in radians. Recall that θ then represents the length of the arc on the unit circle, of which the total length is 2π. Then it seems apparent that, as θ decreases, the arc length coincides with $\sin(\theta)$.

From (2.51) and (2.48a),

$$\begin{aligned}\sin(x-y)+\sin(x+y) &= \sin x\cos y-\cos x\sin y \;+\; \sin x\cos y+\cos x\sin y\\ &= 2\sin x\cos y.\end{aligned}$$

Now let $b=x+y$ and $c=y-x$, so that $x=(b-c)/2$ and $y=(b+c)/2$. It follows that

$$\sin(b)-\sin(c)=2\sin\left(\frac{b-c}{2}\right)\cos\left(\frac{b+c}{2}\right), \tag{2.59}$$

which is one of the sum-to-product identities.

Finally, let $f(x)=\tan(x):=\sin(x)\,/\,\cos(x)$ so that

$$f'(x)=\frac{\cos(x)\cos(x)-\sin(x)(-\sin(x))}{\cos^2(x)}=1+\frac{\sin^2(x)}{\cos^2(x)}=1+\tan^2(x) \tag{2.60}$$

from the quotient rule.

Example 2.6 *To find the derivative of* $y=\arcsin x$*, note that* $\sin y=x$*, so that* y *can be considered an acute angle in a right triangle with a sine ratio of* $x/1$*; see Figure 2.*

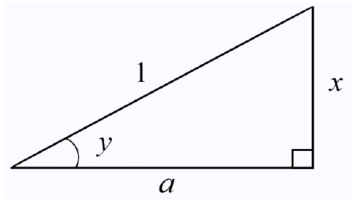


Figure 2: $x=\sin y$ and $\cos y=a$

Differentiating $\sin y=x$ *with respect to* x *and using the chain rule and (2.55) gives*

$$\cos y\cdot\frac{dy}{dx}=1,\quad \text{or}\quad \frac{dy}{dx}=\frac{1}{\cos y}.$$

Note from Figure 2 that $\cos y=a$*. From Pythagoras,* $a^2+x^2=1^2$ *or* $a=\sqrt{1-x^2}$*, so that*

$$\frac{d}{dx}(\arcsin x)=\frac{1}{\sqrt{1-x^2}}.\qquad\blacksquare$$

Example 2.7 *Let* $f(x)=\sin(x)$ *for* $-\pi/2<x<\pi/2$*, with derivative* $f'(x)=\cos x$ *from (2.55). From (2.45) and relation* $\cos^2(x)+\sin^2(x)=1$*, the inverse function* $g(y)=\arcsin(y)$ *has derivative*

$$g'(y)=\frac{1}{\cos(\arcsin(y))}=\frac{1}{\sqrt{1-\sin^2(\arcsin(y))}}=\frac{1}{\sqrt{1-y^2}},$$

which agrees with Example 2.6. Similarly, let $f(x)=\tan(x)$*, for* $-\pi/2<x<\pi/2$ *with inverse function* $g(y)=\arctan(y)$*, so that, from (2.60),*

$$g'(y)=\frac{1}{1+\tan^2(\arctan(y))}=\frac{1}{1+y^2}. \tag{2.61}$$

Now let z *be a constant. Using (2.61) and the chain rule gives*

$$\frac{d}{dx}\arctan(z-x)=-\frac{1}{1+(z-x)^2}, \tag{2.62}$$

which we will use below in Example 2.52. $\blacksquare$

2.2.3 Mean Value Theorem and Function Extreme Points

Theorem (Mean Value Theorem, MVT): Let f be a continuous function on its domain $[a, b]$, $b > a$, and differentiable on (a, b). Then $\exists \xi \in (a, b)$ such that $f(b) - f(a) = f'(\xi)(b - a)$. The MVT is perhaps more easily remembered as

$$\frac{f(b) - f(a)}{b - a} = f'(\xi), \quad b - a \neq 0. \tag{2.63}$$

The proof is given below, after we prove Rolle's theorem. Many common and important calculus results (see the list below) hinge on this result, or a generalization of it, and we will also see it used extensively in the multivariate setting. The MVT becomes intuitive from Figure 3, for a differentiable (and, thus, continuous) function, and such that f is continuous (from the right, and the left, respectively) also at endpoints a and b, as stated in the theorem.

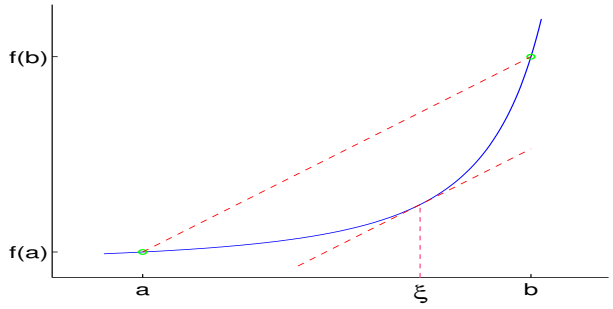


Figure 3: The mean value theorem of the differential calculus

Still, without a proof, a convincing argument and clever graphic are not adequate. Pugh (Real Mathematical Analysis, 2nd ed., 2015, p. 4) says it quite well, and is the only analysis book I have ever seen that discusses this in such detail. Here is an excerpt:

> When is a mathematical statement accepted as true? Generally, mathematicians would answer "Only when it has a proof inside a familiar mathematical framework." A picture may be vital in getting you to believe a statement. An analogy with something you know to be true may help you understand it. An authoritative teacher may force you to parrot it. A formal proof, however, is the ultimate and only reason to accept a mathematical statement as true.
>
> There has been a tendency in recent years to take the notion of proof down from its pedestal. Critics point out that standards of rigor change from century to century. New gray areas appear all the time. Is a proof by computer an acceptable proof? Is a proof that is spread over many journals and thousands of pages, that is too long for any one person to master, a proof? And of course, venerable Euclid is full of flaws, some filled in by Hilbert, others possibly still lurking.
>
> Clearly it is worth examining closely and critically the most basic notion of mathematics, that of proof. On the other hand, it is important to bear in mind that all distinctions and niceties about what precisely constitutes a proof are mere quibbles compared to the enormous gap between any generally accepted version of a proof and the notion of a convincing argument. Compare Euclid, with all his flaws to the most eminent of the ancient exponents of the convincing argument – Aristotle. Much of Aristotle's reasoning was brilliant, and he certainly convinced most thoughtful people for over a thousand years. In some cases his analyses were exactly right, but in others, such as heavy objects falling faster than light ones, they turned out to be totally wrong. In contrast, there is not to my knowledge a single theorem stated in Euclid's Elements that in the course of two thousand years turned out to be false. That is quite an astonishing record, and an extraordinary validation of proof over convincing argument.

Theorem (Rolle): Suppose that f is a function defined and continuous on an interval $[a,b]$, that it is differentiable in (a,b), and that $f(a)=f(b)$. Then

$$\exists c \in (a,b) \text{ such that } f'(c)=0. \tag{2.64}$$

The equivalent contrapositive will be used below:

$$\nexists c \in (a,b) \text{ such that } f'(c)=0 \;\Rightarrow\; f(a)\neq f(b). \tag{2.65}$$

Proof: We start with the EVT (2.29), which guarantees that f attains its largest value M and its smallest value m on $[a,b]$. There are two possibilities: either $M=m$ or $M>m$. In the former case, the inequality $m\leq f(x)\leq M$ implies that f is constant on $[a,b]$, so $f'(x)=0$ for all $x\in(a,b)$ and we can take for c any point in (a,b). If $M>m$, the assumption that $f(a)=f(b)$ shows that at least one of M and m is attained at a point $c\in(a,b)$. By Fermat's Theorem (2.43), $f'(c)=0$.

Proof of the MVT:

The function

$$F(x)=f(x)-\frac{f(b)-f(a)}{b-a}x$$

satisfies $F(a)=F(b)$. Since linear functions are differentiable (and, hence, continuous), F satisfies all the hypotheses of Rolle's Theorem (2.64). It follows that there exists $c\in(a,b)$ such that $F'(c)=0$. Clearly,

$$F'(x)=f'(x)-\frac{f(b)-f(a)}{b-a}$$

so

$$0=F'(c)=f'(c)-\frac{f(b)-f(a)}{b-a}.$$

Example 2.8 *As in Pons, Thm 5.4.5, let f and g be functions, differentiable on $(0,\infty)$ and continuous on $[0,\infty)$. Prove: If $f'(x)\leq g'(x)$ for every $x\in(0,\infty)$ and $f(0)=g(0)$, then $f(x)\leq g(x)$ for every $x\in[0,\infty)$.*

Proof: Let $h=g-f$, so $h'(x)=g'(x)-f'(x)\geq 0$ for all $x\in(0,\infty)$. Fix $x_0\in(0,\infty)$ and apply the MVT to $h\in(0,x_0)$, showing $\exists c\in(0,x_0)$ such that

$$h'(c)=\frac{h(x_0)-h(0)}{x_0-0}.$$

The quotient and the denominator are nonnegative; thus it must be the case that $h(x_0)-h(0)\geq 0$. Substituting for f and g,

$$0\leq h(x_0)-h(0)=g(x_0)-f(x_0)-(g(0)-f(0))=g(x_0)-f(x_0),$$

implying $f(x_0)\leq g(x_0)$. ■

We now collect some further useful results.

- If $\forall x\in I$, $f'(x)=0$, then the MVT (2.63) implies that, $\forall x,y\in I$, $f(y)=f(x)$, i.e.,

$$f \textbf{ is constant on } I. \tag{2.66}$$

- Let I be the open interval (a, b). If f is differentiable on I and, $\forall x \in I$, $|f'(x)| \leq M$, then the MVT implies that $|f(y) - f(x)| \leq M|y - x|$ for all $x, y \in I$. This is referred to as the *(global) Lipschitz condition*.

- The MVT is mainly used for proving other results, including the fundamental theorem of calculus (see §2.4.2), the validity of interchanging derivative and integral (§5.3), and the fundamental optimization results, proven below, in (2.74) and (2.75).

- If (i) $f'(c) > 0$ for some point $c \in I$, and (ii) f' is continuous at c, then, from (2.23), $\exists\, \delta > 0$ such that, $\forall x \in (c - \delta, c + \delta)$, $f'(x) > 0$, i.e., f is increasing on that interval. See (2.69) below for proof. Condition (ii) cannot be dropped: See, e.g., Stoll, 2021, the remark on p. 199.

- The MVT can be generalized to the *Cauchy* or *Ratio Mean Value Theorem*, as stated and proved below. It is used, for example, to rigorously prove l'Hôpital's rule (2.44). Given the prominence of the MVT, Stoll (2001, p. 204) argues that it could justifiably called the *Fundamental Theorem of Differential Calculus*.

Theorem (Cauchy Mean Value Theorem): If f and g are continuous functions on $[a, b]$ and differentiable on $I = (a, b)$, then $\exists\, c \in I$ such that $[f(b) - f(a)]g'(c) = [g(b) - g(a)]f'(c)$ or, easier to remember, if $g(b) - g(a) \neq 0$,

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(c)}{g'(c)}. \tag{2.67}$$

Notice that this reduces to the usual mean value theorem when $g(x) = x$.

> Proof: Let $h(x) = [f(b) - f(a)]g(x) - [g(b) - g(a)]f(x)$. Then, from (2.20) and (2.21), h is continuous on $[a, b]$; and, from (2.37) and (2.38), differentiable on (a, b) with
>
> $$h(a) = f(b)g(a) - f(a)g(b) = h(b).$$
>
> Thus by Rolle's theorem, there exists $c \in (a, b)$ such that $h'(c) = 0$, which gives the result.
>
> If $g'(x) \neq 0$ for all $x \in (a, b)$, then (2.65) implies $g(a) \neq g(b)$, so that (2.67) can be written as
>
> $$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(c)}{g'(c)}.$$

Remark: Recall the intermediate value theorem (IVT) in (2.30). Let $I \subset \mathbb{R}$ be an interval and let $f : I \to \mathbb{R}$ be differentiable on I. If f' is continuous on I, then the IVT applied to f' implies that, for $a, b \in I$ with $a < b$, $\alpha = f'(a)$, $\beta = f'(b)$ and a value $\gamma \in \mathbb{R}$ with either $\alpha < \gamma < \beta$ or $\alpha > \gamma > \beta$, $\exists\, c \in (a, b)$ such that $f'(c) = \gamma$. More intriguing is the fact that this still holds *even if f' is not continuous*, a result attributed to Jean Gaston Darboux (1842–1917); see Stoll (2001, p. 184), Browder (1996, Thm. 4.25) or Pugh (2002, p. 144). It is referred to as the *Intermediate Value Theorem for Derivatives*.

> **Remark**: The need occasionally arises to construct simple graphics like Figure 3, and it is often expedient to use the plotting and graphics generation capabilities of Matlab or other such software. In this case, the graph was constructed using the function $f(x) = 1/(1 - x)^2$ with endpoints $a = 0.6$ and $b = 0.9$.

This is also a good excuse to illustrate Matlab's symbolic toolbox (which uses the Maple computing engine). The top third of the code in Listing 2 uses some basic commands from the symbolic toolbox to compute ξ based on our choice of f, a and b. The rest of the code constructs Figure 3.

While Matlab supports interactive graphics editing, use of the native graphics commands ("batch code") in Listing 2 is not only faster the first time around (once you are familiar with them of course), but ensures that the picture can be identically reproduced. ■

We now turn to the most basic concepts of function minimization / maximization. The first theorem is the same as Fermat's theorem given above in (2.43), which we needed for proving Rolle.

Theorem: Let I be a neighborhood of x_0 and suppose that the function $f : I \to \mathbb{R}$ is differentiable at x_0. If the point x_0 is either a maximizer or a minimizer of the function $f : I \to \mathbb{R}$, then

$$f'(x_0) = 0. \tag{2.68}$$

Proof: Observe that, by the definition of a derivative,

$$\lim_{x \to x_0, x < x_0} \frac{f(x) - f(x_0)}{x - x_0} = \lim_{x \to x_0, x > x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0).$$

First suppose that x_0 is a maximizer. Then

$$\frac{f(x) - f(x_0)}{x - x_0} \geq 0 \quad \text{for } x \text{ in } I \text{ with } x < x_0,$$

and hence, from (2.15),

$$f'(x_0) = \lim_{x \to x_0, x < x_0} \frac{f(x) - f(x_0)}{x - x_0} \geq 0.$$

On the other hand,

$$\frac{f(x) - f(x_0)}{x - x_0} \leq 0 \quad \text{for } x \text{ in } I \text{ with } x > x_0,$$

and hence

$$f'(x_0) = \lim_{x \to x_0, x > x_0} \frac{f(x) - f(x_0)}{x - x_0} \leq 0.$$

Thus, $f'(x_0) = 0$.

In the case where x_0 is a minimizer, the same proof applies, with inequalities reversed.

Theorem: Let I be an open interval and the function $f : I \to \mathbb{R}$ be differentiable. Suppose that $f'(x) > 0$ for all x in I. Then

$$f : I \to \mathbb{R} \textbf{ is strictly increasing.} \tag{2.69}$$

```
function meanvaluetheorem
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%         Use the symbolic toolbox to compute xi           %%%%%
 syms x xi real  % declare x and xi to be real symbolic variables  %
 f=1/(1-x)^2     % our function                                     %
 a=0.6; b=0.9;   % use these two end points                        %
 fa=subs(f,'x',a); fb=subs(f,'x',b); % evaluate f at a and b       %
 ratio=(fb-fa)/(b-a)  % slope of the line                         %
 df=diff(f)           % first derivative of f                     %
 xi = solve(df-ratio) % find x such that f'(x) = ratio            %
 xi=eval(xi(1))  % there is only one real solution               %
 subs(df,'x',xi) % just check if equals ratio                     %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Plot function and the slope line
xx=0.57:0.002:0.908; ff=1./(1-xx).^2; h=plot(xx,ff)
hold on
  h=plot([a ; b],[fa ; fb],'go'); set(h,'linewidth',28)
hold off
set(h,'LineWidth',1.5), bot=-6; axis([0.53 0.96 bot 125])
set(gca,'fontsize',21,'Box','off', ...
        'YTick',[fa fb], 'YtickLabel',{'f(a)' ; 'f(b)'}, ...
        'XTick',[a b],   'XTickLabel',{'a' ; 'b'})
h=line([a b],[fa fb]);
set(h,'linestyle','--','color',[1 0 0],'linewidth',0.8)

% plot line y-y0 = m(x-x0) where m is slope and goes through (x0,y0)
x0=xi; y0=subs(f,'x',x0);
xa=a+0.4*(b-a);        xb=b-0.0*(b-a);
ya = y0+ratio*(xa-x0); yb = y0+ratio*(xb-x0);
h=line([xa xb],[ya yb]);
set(h,'linestyle','--','color',[1 0 0],'linewidth',0.8)

% vertical line at xi with label at xi
h=line([xi xi],[bot y0]);
set(h,'linestyle','--','color',[1 0.4 0.6],'linewidth',1.2)
text(xi-0.005,-13,'\xi','fontsize',24)
% Text command (but not the XTickLabel) supports use of
%  LaTeX-like text strings
```

Program Listing 2: Computes ξ in the mean value theorem, and creates Figure 3

Proof: Let u and v be points in I with $u < v$. Then we can apply the Mean Value Theorem to the restriction of f to the closed bounded interval $[u, v]$ and choose a point x_0 in the open interval (u, v) at which

$$f'(x_0) = \frac{f(v) - f(u)}{v - u}.$$

Since $f'(x_0) > 0$ and $v - u > 0$, it follows that $f(u) < f(v)$.

By replacing $f : I \to \mathbb{R}$ with $-f : I \to \mathbb{R}$, the above implies that if $f : I \to \mathbb{R}$ has a negative derivative at each point x in I, then $f : I \to \mathbb{R}$ is strictly decreasing.

Definition: A point x_0 in the domain of a function $f : D \to \mathbb{R}$ is said to be a local maximizer for f provided that there is some $\delta > 0$ such that

$$f(x) \leq f(x_0) \quad \text{for all } x \text{ in } D \text{ such that } |x - x_0| < \delta.$$

We call x_0 a local minimizer for f provided that there is some $\delta > 0$ such that

$$f(x) \geq f(x_0) \quad \text{for all } x \text{ in } D \text{ such that } |x - x_0| < \delta.$$

The above result (2.68) asserts that, if I is a neighborhood of x_0 and $f : I \to \mathbb{R}$ is differentiable at x_0, then for x_0 to be either a local minimizer or a local maximizer for f, it is necessary that

$$f'(x_0) = 0.$$

However, knowing that $f'(x_0) = 0$ does not guarantee that x_0 is either a local maximizer or a local minimizer. For instance, if $f(x) = x^3$ for all x, then $f'(0) = 0$, but the point 0 is neither a local maximizer nor a local minimizer for the function f. In order to establish criteria that are sufficient for the existence of local maximizers and local minimizers, it is necessary to introduce higher derivatives.

The second derivative of f, if it exists, is the derivative of f', and denoted f'' or $f^{(2)}$, and likewise for higher order derivatives. If $f^{(r)}$ exists, then f is said to be rth *order differentiable*, and if $f^{(r)}$ exists for all $r \in \mathbb{N}$, then f is *infinitely differentiable*, or *smooth* (see, e.g., Pugh, 2002, p. 147). Let $f^{(0)} \equiv f$. As differentiability implies continuity, it follows that, if f is rth order differentiable, then $f^{(r-1)}$ is continuous, and that smooth functions and all their derivatives are continuous. If f is rth order differentiable and $f^{(r)}$ is continuous, then f is *continuously* rth *order differentiable*, and f is of *class* $\mathcal{C}^r$. An infinitely differentiable function is of *class* $\mathcal{C}^\infty$.

For a differentiable function $f : I \to \mathbb{R}$ that has as its domain an open interval I, we say that $f : I \to \mathbb{R}$ has one derivative if $f : I \to \mathbb{R}$ is differentiable and define $f^{(1)}(x) = f'(x)$ for all x in I. If the function $f' : I \to \mathbb{R}$ itself has a derivative, we say that $f : I \to \mathbb{R}$ has two derivatives, or has a second derivative, and denote the derivative of $f' : I \to \mathbb{R}$ by $f'' : I \to \mathbb{R}$ or by $f^{(2)} : I \to \mathbb{R}$. Now let k be a natural number for which we have defined what it means for $f : I \to \mathbb{R}$ to have k derivatives and have defined $f^{(k)} : I \to \mathbb{R}$. Then $f : I \to \mathbb{R}$ is said to have $k + 1$ derivatives if $f^{(k)} : I \to \mathbb{R}$ is differentiable, and we define $f^{(k+1)} : I \to \mathbb{R}$ to be the derivative of $f^{(k)} : I \to \mathbb{R}$. In this context, is it useful to denote $f(x)$ by $f^{(0)}(x)$.

In general, if a function has k derivatives, it does not necessarily have $k + 1$ derivatives. For instance, the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = |x|x$ for all x is differentiable but does not have a second derivative.

The goal now is to determine what conditions on the second derivative of function f are required in order to conclude that $f'(x_0) = 0 \Rightarrow x_0$ is a local minimizer or maximizer of f. The conditions are given below in (2.74) and (2.75). To prove them, we first require a preliminary result.

Lemma: For an open interval $I \subset \mathbb{R}$, let $g : I \to \mathbb{R}$ be differentiable, and let $x_0 \in I$. If $g'(x_0) > 0$, then

$$\exists \delta > 0 \text{ such that } 0 < |x - x_0| < \delta \;\; \Rightarrow \;\; [g(x) - g(x_0)] / [x - x_0] > 0. \tag{2.70}$$

Proof: By definition of the derivative, and the assumption $g'(x_0) > 0$,

$$g'(x_0) = \lim_{x \to x_0} \frac{g(x) - g(x_0)}{x - x_0} > 0.$$

If g' is continuous at x_0, then (2.70) follows from (2.23) applied to g'. Now consider the case without the continuity assumption. Define $h : I \to \mathbb{R}$ as

$$h(x) = \begin{cases} [g(x) - g(x_0)] / [x - x_0], & \text{if } x \neq x_0, \\ g'(x_0), & \text{if } x = x_0. \end{cases}$$

As g is differentiable, g is continuous from (2.42). From (2.20) and (2.22), h is continuous for $x \neq x_0$. As $\lim_{x \to x_0} h(x) = h(x_0)$, (2.19) implies h is continuous also at $x = x_0$, and thus $h : I \to \mathbb{R}$ is continuous. Result (2.70) now follows from (2.23) applied to h.

Before proceeding, we work a bit further with this lemma. It is equivalent to the remark in Stoll (2021, p. 199):

> It needs to be emphasized that if the derivative of a function f is positive at a point c, then this does not imply that f is increasing on an interval containing c; it could be non-monotone on any interval containing c. If $f'(c) > 0$, the only conclusion that can be reached is: $\exists \delta > 0$ such that
>
> $$\forall x \in (c - \delta, c),\ f(x) < f(c) \text{ and } \forall x \in (c, c + \delta),\ f(x) > f(c). \tag{2.71}$$
>
> This does not mean that f is increasing on $(c - \delta, c + \delta)$.

While perhaps a bit tricky to visualize because f' is not continuous, imagine, for example, a differentiable (and thus continuous) function that is (pathologically) oscillatory for $x \in (c - \delta, c + \delta)$ but such that (2.71) is satisfied, i.e., all its values for $x \in (c - \delta, c)$ lie below $f(c)$, and all its values are $x \in (c, c + \delta)$ lie above $f(c)$. Here is the proof of (2.71):

Proof: By hypothesis, $f'(c)$ exists and $f'(c) > 0$. Let $\epsilon = f'(c) > 0$. Then by existence of the derivative, $\exists \delta > 0$ such that

$$\left| \frac{f(x) - f(c)}{x - c} - f'(c) \right| < \epsilon, \;\; \forall x \in I \text{ with } 0 < d(x, c) < \delta. \tag{2.72}$$

This implies that

$$\underbrace{f'(c) - \epsilon}_{=0} < \frac{f(x) - f(c)}{x - c}, \;\; \forall x \in I \text{ with } 0 < d(x, c) < \delta. \tag{2.73}$$

Note that, whenever $x > c$ (with $0 < x - c < \delta$), the denominator is positive, so that $f(x) > f(c)$. Similarly, whenever $c > x$ (with $0 < c - x < \delta$), we must have $f(c) > f(x)$. We have therefore proven the existence of a δ as required.

Theorem: Let I be an open interval containing the point x_0 and suppose that the function $f : I \to \mathbb{R}$ has a second derivative. Suppose that $f'(x_0) = 0$. Then,

$$\textbf{If } f''(x_0) > 0\textbf{, then } x_0 \textbf{ is a local minimizer of } f. \tag{2.74}$$

$$\textbf{If } f''(x_0) < 0\textbf{, then } x_0 \textbf{ is a local maximizer of } f. \tag{2.75}$$

Proof: First suppose that $f''(x_0) > 0$ (and not necessarily continuous). Since

$$f''(x_0) = \lim_{x \to x_0} \frac{f'(x) - f'(x_0)}{x - x_0} > 0,$$

it follows from Lemma (2.70) that there is a $\delta > 0$ such that the open interval $(x_0 - \delta, x_0 + \delta)$ is contained in I and

$$\frac{f'(x) - f'(x_0)}{x - x_0} > 0 \quad \text{if } x \text{ belongs to } (x_0 - \delta, x_0 + \delta). \tag{2.76}$$

But $f'(x_0) = 0$, so (2.76) amounts to the assertion that

$$f'(x) > 0 \text{ if } x_0 < x < x_0 + \delta \quad \text{and} \quad f'(x) < 0 \text{ if } x_0 - \delta < x < x_0. \tag{2.77}$$

From the first inequality in (2.77), the MVT implies $\exists \xi \in (x_0, x)$ such that

$$\frac{f(x) - f(x_0)}{x - x_0} = f'(\xi) > 0 \;\Rightarrow\; f(x) > f(x_0);$$

while from the second inequality in (2.77), the MVT implies $\exists \xi \in (x, x_0)$ such that

$$\frac{f(x_0) - f(x)}{x_0 - x} = f'(\xi) < 0 \;\Rightarrow\; f(x) > f(x_0).$$

Thus, for $0 < |x - x_0| < \delta$, $f(x) > f(x_0)$, which is (2.74).

A similar argument applies for $f''(x_0) < 0$ to prove (2.75).

The preceding theorem provides no information about $f(x_0)$ as a local extreme point if both $f'(x_0) = 0$ and $f''(x_0) = 0$. As we see from examining functions of the form $f(x) = cx^n$ for all x at $x_0 = 0$, if $f'(x_0) = 0$ and $f''(x_0) = 0$, then x_0 may be a local maximizer, a local minimizer, or neither.

2.2.4 Exponential and Logarithm

> For the sake of brevity, we will always represent this number 2.718281828459... by the letter e. (Leonhard Euler)

The *exponential function* arises ubiquitously in mathematics, and so it is worth spending some time understanding it. As in Lang (1997, §4.1, §4.2), let $f : \mathbb{R} \to \mathbb{R}$ be such that

$$\text{(i) } f'(x) = f(x) \quad \text{and} \quad \text{(ii) } f(0) = 1. \tag{2.78}$$

From the product rule (2.38), the chain rule (2.40), and using (i),

$$[f(x) f(-x)]' = -f(x) f'(-x) + f(-x) f'(x) = -f(x) f(-x) + f(-x) f(x) = 0,$$

so that, from (2.66), $f(x) f(-x)$ is constant, and from (ii), equals 1. Thus, $f(x) \neq 0$ and $f(-x) = 1/f(x)$. As f is differentiable, f is continuous. From (ii) and the contrapositive of the IVT (2.30), $\forall x$, $f(x) > 0$. Thus, from (i), $\forall x$, $f'(x) > 0$, i.e., f is strictly increasing. Further, as $f'' = f' = f$, f is strictly convex.

Theorem:

A function f satisfying (i) and (ii) in (2.78) is unique. (2.79)

Proof: Suppose g is any function such that $g' = g$. From (2.39), differentiating g/f (and $\forall x$, $f(x) \neq 0$) yields 0. Hence, from (2.66), $g/f = K$ for some constant K, and thus $g = Kf$. If $g(0) = 1$, then $g(0) = Kf(0)$ so that $K = 1$ and $g = f$.

Theorem:

$$f(x+y) = f(x) f(y) \quad \text{and} \quad f(nx) = [f(x)]^n, \quad n \in \mathbb{N}. \tag{2.80}$$

Proof: For the first result in (2.80), fix a number a, and consider the function $g(x) = f(a+x)$. Then $g'(x) = f'(a+x) = f(a+x) = g(x)$. From the previous uniqueness proof, $g' = g$ implies $g(x) = Kf(x)$ for some constant K. Letting $x = 0$ shows that $K = g(0) = f(a)$. Hence, $f(a+x) = f(a)f(x)$ for all x, as contended. For the second result in (2.80), this is true when $n = 1$, and assuming it for n, we have

$$f((n+1)a) = f(na+a) = f(na)f(a) = f(a)^n f(a) = f(a)^{n+1},$$

by induction. The second result in (2.80) also holds for $n \in \mathbb{R}$.

Function f is the exponential, written $\exp(\cdot)$. Defining $e = f(1)$, and using that the second result in (2.80) also holds for $n \in \mathbb{R}$, we can write

$$\exp(x) = f(x) = f(1 \cdot x) = [f(1)]^x = e^x. \tag{2.81}$$

As shown above, $f(x)$ is strictly increasing; and, as $f(0) = 1$, we have that $f(1) = e > 1$.

It follows from (1.22) and the fact that $e > 1$ that

$$\forall k \in \mathbb{N}, \quad \lim_{n\to\infty} e^n/n^k = \infty, \quad n \in \mathbb{N}. \tag{2.82}$$

Now replace n by x, for $x \in \mathbb{R}_{>0}$. Use of the quotient rule (2.39) gives

$$\frac{d}{dx}\left(\frac{e^x}{x^k}\right) = \frac{x^k e^x - e^x k x^{k-1}}{x^{2k}} = \frac{e^x}{x^k}(1 - k/x),$$

which is positive for $k < x$, i.e., for x large enough, e^x/x^k is increasing. This and the limit result for $n \in \mathbb{N}$ implies that

$$\lim_{x\to\infty} \frac{e^x}{x^k} = \infty, \quad \text{for all } k \in \mathbb{N}, \tag{2.83}$$

a result we will use below in (2.102). Recall the discussion just above (2.45): As $f(x)$ is strictly increasing, the inverse function g exists; and as $f(x) > 0$, $g(y)$ is defined for $y > 0$. From (ii), $g(1) = 0$. From (2.45) and (i) in (2.78),

$$g'(y) = \frac{1}{f'(g(y))} = \frac{1}{f(g(y))} = \frac{1}{y}. \tag{2.84}$$

For $a > 0$, the weighted sum and chain rules for differentiation, (2.41) and (2.40), yield

$$[g(ax) - g(x)]' = ag'(ax) - g'(x) = \frac{a}{ax} - \frac{1}{x} = 0,$$

and (2.66) then implies $g(ax) - g(x) = c$ or $g(ax) = c + g(x)$. Letting $x = 1$ gives $g(a) = c + g(1) = c$, which then implies

$$g(ax) = g(a) + g(x). \tag{2.85}$$

By induction, $g(x^n) = ng(x)$. Function g is the *natural logarithm*, denoted $\log(y)$, $\log y$ or, from the French *logarithm natural*, $\ln y$. Thus,

$$\ln 1 = 0, \tag{2.86}$$

$$\ln x^n = n \ln x, \quad n \in \mathbb{N}, \ x > 0, \tag{2.87}$$

and, from (2.84),

$$\frac{d}{dx} \ln x = \frac{1}{x}, \quad x > 0. \tag{2.88}$$

As $\ln 1 = 0$, write $0 = \ln 1 = \ln(x/x) = \ln x + \ln(1/x)$ from (2.85), so that $\ln x^{-1} = -\ln x$. The last two results generalize to (see Stoll, 2001, p. 234)

$$\ln(x^p) = p \cdot \ln(x), \quad p \in \mathbb{R}, \ x \in \mathbb{R}_{>0}. \tag{2.89}$$

The reader can have a peak at Example 2.41 below regarding (2.87) and (2.89), where the log function is defined in terms of an integral. Based on their properties, the exponential and logarithmic functions are also used in the following way:

$$\textbf{For } r \in \mathbb{R} \textbf{ and } x \in \mathbb{R}_{>0}, \ x^r \textbf{ is defined by } x^r := \exp(r \ln x). \tag{2.90}$$

Example 2.9 *To evaluate* $\lim_{x \to 0^+} x^x$*, use l'Hôpital's rule and* (2.88) *to see that*

$$\lim_{x \to 0^+} x \ln x = \lim_{x \to 0^+} \frac{\ln x}{1/x} = \lim_{x \to 0^+} \frac{1/x}{-x^{-2}} = -\lim_{x \to 0^+} x = 0.$$

Then, by (2.19) *and the continuity of the exponential function,*

$$\lim_{x \to 0^+} x^x = \lim_{x \to 0^+} \exp(\ln x^x) = \lim_{x \to 0^+} \exp(x \ln x) = \exp\left(\lim_{x \to 0^+} x \ln x\right) = \exp 0 = 1. \quad \blacksquare$$

Example 2.10 *We will make important use of the following two basic limit results. From the continuity of the exponential function and use of l'Hôpital's rule,*

$$\lim_{k \to \infty} k^{1/k} = \lim_{k \to \infty} \exp\left(\ln k^{1/k}\right) = \lim_{k \to \infty} \exp\left(\frac{\ln k}{k}\right) = \exp \lim_{k \to \infty} \frac{\ln k}{k} = \exp \lim_{k \to \infty} (1/k) = 1.$$

Similarly, and again using the continuity of the exponential function, for any $a \in \mathbb{R}_{>0}$,

$$\lim_{n\to\infty} \sqrt[n]{a} = \lim_{n\to\infty} a^{1/n} = \lim_{n\to\infty} \exp\left(\frac{\ln a}{n}\right) = \exp \lim_{n\to\infty}\left(\frac{\ln a}{n}\right) = \exp(0) = 1. \tag{2.91}$$

We now give a proof of (2.91) *using much less sophisticated machinery. As in Petrovic (Example 2.9.1), first let* $a \geq 1$. *Recall Bernoulli's inequality (which, for* $x \geq 0$, *is just the first term in the binomial theorem expansion* (1.21)*; or can be proven by induction): For* $x > -1$ *and* $n \in \mathbb{N}$, $(1+x)^n \geq 1 + nx$. *With* $x := \sqrt[n]{a} - 1 \geq 0$,

$$a = (1+x)^n \geq 1 + nx = 1 + n(\sqrt[n]{a} - 1), \quad \text{or} \quad 0 \leq \sqrt[n]{a} - 1 \leq \frac{a-1}{n}.$$

Taking the limit as $n \to \infty$ *and use of the Squeeze Theorem* (2.3) *implies* $\lim a_n = 1$.

Now consider the case for which $0 < a < 1$. *Let* $b = 1/a > 1$. *From the previous result,* $1 = \lim \sqrt[n]{b} = \lim b^{1/n}$. *From the limits of ratios result* (2.14),

$$\lim \sqrt[n]{a} = \lim \frac{1}{\sqrt[n]{b}} = \frac{\lim 1}{\lim \sqrt[n]{b}} = \frac{1}{1} = 1.$$

(Enter a positive number in your calculator, repeatedly press the $\sqrt{}$ *key, and see what happens: Either the key will break, or a 1 will result).* ■

Example 2.11 *Let* $f(x) = x^r$, *for* $r \in \mathbb{R}$ *and* $x \in \mathbb{R}_{>0}$. *From (2.90) and the chain rule,*

$$f'(x) = \exp(r \ln x)\frac{r}{x} = x^r \frac{r}{x} = r x^{r-1},$$

which extends the results in Example 2.2 in a natural way. ■

Example 2.12 *Consider the case when the variable is not the base, but the exponent:*

$$\textbf{\textit{For}}\ t \in \mathbb{R}_{>0}\ \textbf{\textit{and}}\ f(x) = t^x\textbf{\textit{,}}\ f'(x) = t^x \ln t\textbf{\textit{.}} \tag{2.92}$$

From (2.90), $f(x) = \exp(x \ln t)$. *Then* $f'(x) = \exp(x \ln t)(\ln t) = t^x \ln t$ *(chain rule).* ■

This next example gives an application of (2.92). We need the following defintion, which we take from §2.5.5, where further detail will be found. Let $\{f_n(x)\}$ be a *sequence of functions* with the same domain, say D. The function f is the *pointwise limit* of sequence $\{f_n\}$, or $\{f_n\}$ *converges pointwise* to f, if, $\forall x \in D$, $\lim_{n\to\infty} f_n(x) = f(x)$. That is, $\forall x \in D$ and for every given $\epsilon > 0$, $\exists N \in \mathbb{N}$ such that $|f_n(x) - f(x)| < \epsilon$, $\forall n > N$.

Example 2.13 *(Stade, Fourier Analysis, p. 157) For domain* $D = [0, 2\pi]$, $N \in \mathbb{N}$, *let* $f_N : D \to \mathbb{R}$ *be the function defined by*

$$f_N(x) = N\left(\frac{x}{2\pi}\right)^N \sqrt{2\pi - x}. \tag{2.93}$$

Also let $f(x) = 0$ *for all* $x \in [0, 2\pi]$. *Show that the* f_N*'s converge pointwise to* f *but do not converge to* f *in norm.*

Solution: Let's first take care of the cases $x = 0$ *and* $x = 2\pi$, *which are easy:* $f_N(0) = f_N(2\pi) = 0 \to 0 = f(0) = f(2\pi)$ *as* $N \to \infty$, *as required.*

Next, for any fixed element x of $(0, 2\pi)$, use l'Hôpital's rule and (2.92) as follows:

$$\begin{aligned}\lim_{N\to\infty} f_N(x) &= \lim_{N\to\infty} N\left(\frac{x}{2\pi}\right)^N \sqrt{2\pi - x} = \sqrt{2\pi - x} \lim_{N\to\infty} \frac{N}{(2\pi/x)^N} \\ &= \sqrt{2\pi - x} \lim_{N\to\infty} \frac{1}{(2\pi/x)^N \ln(2\pi/x)} \\ &= \frac{\sqrt{2\pi - x}}{\ln(2\pi/x)} \lim_{N\to\infty} \left(\frac{x}{2\pi}\right)^N = 0.\end{aligned}$$

(The limit on the right is zero because $0 < x < 2\pi$.) So the f_N 's converge pointwise to f on $[0, 2\pi]$, as required. ■

Example 2.14 *For $x > 0$ and $p \in \mathbb{R} \setminus \{0\}$, the chain rule and (2.88) imply*

$$\frac{d}{dx} (\ln x)^p = \frac{p (\ln x)^{p-1}}{x}, \tag{2.94}$$

so that, dividing both sides by p, integrating both sides (and using the fundamental theorem of calculus; see §2.4.2 below),

$$\frac{(\ln x)^p}{p} = \int \frac{dx}{x (\ln x)^{1-p}}, \tag{2.95}$$

which we require below in Example 2.62. Also, from (2.88) and the chain rule,

$$\frac{d}{dx} \ln (\ln x) = \frac{1}{\ln x} \frac{d}{dx} \ln x = \frac{1}{x \ln x}, \tag{2.96}$$

also required in Example 2.62. ■

Example 2.15 *For $y > 0$, $k \in \mathbb{R}$, and $f : \mathbb{R} \to \mathbb{R}$ given by $f(p) = y^{kp}$,*

$$f'(p) = \frac{d}{dp} y^{kp} = \frac{d}{dp} \exp (kp \ln y) = \exp (kp \ln y)\, k \ln y = y^{kp} k \ln y. \tag{2.97}$$

With $k = -1$, $x > 1$, and $y = \ln x$, (2.97) implies

$$\frac{d}{dp} \left((\ln x)^{-p}\right) = - (\ln x)^{-p} \ln (\ln x).$$

Also, for $y > 0$ and $k = -1$, (2.97) implies

$$\frac{d}{dp} y^{1-p} = y \frac{d}{dp} y^{-p} = -y^{1-p} \ln y, \tag{2.98}$$

which we will use in the next example. ■

Example 2.16 *In microeconomics, a* utility function, *$U(\cdot)$, is a preference ordering for different goods of choice ("bundles" of goods and services, amount of money, etc.) For example, if bundle A is preferable to bundle B, then $U(A) > U(B)$. Let $U : A \to \mathbb{R}$, $A \subset \mathbb{R}_{>0}$, be a continuous and twice differentiable utility function giving a preference ordering for overall wealth, W. Not surprisingly, one assumes that $U'(W) > 0$, i.e., people prefer more wealth to less, but also that $U''(W) < 0$, i.e., the more wealth you have, the less additional utility you reap upon obtaining a fixed increase in wealth. (In this case, U is a concave*

function and the person is said to be risk-averse.*) A popular choice of* U *is* $U(W;\gamma) = W^{1-\gamma}/(1-\gamma)$ *for a fixed parameter* $\gamma \in \mathbb{R}_{>0} \setminus \{1\}$ *and* $W > 0$*. (Indeed, an easy calculation verifies that* $U'(W) > 0$ *and* $U''(W) < 0$*). Interest centers on the limit of* U *as* $\gamma \to 1$*. In this case,* $\lim_{\gamma\to 1} W^{1-\gamma} = 1$ *and* $\lim_{\gamma\to 1}(1-\gamma) = 0$ *so that l'Hôpital's rule is not applicable. However, as utility is a relative measure, we can let* $U(W;\gamma) = (W^{1-\gamma}-1)/(1-\gamma)$ *instead. Then, from (2.98),* $(d/d\gamma)\, W^{1-\gamma} = -W^{1-\gamma}\ln W$*, so that*

$$\lim_{\gamma\to 1} U(W;\gamma) = \lim_{\gamma\to 1} \frac{W^{1-\gamma}-1}{1-\gamma} = \lim_{\gamma\to 1} \frac{(d/d\gamma)\,(W^{1-\gamma}-1)}{(d/d\gamma)\,(1-\gamma)} = \lim_{\gamma\to 1} W^{1-\gamma}\ln W = \ln W. \qquad \blacksquare$$

Example 2.17 *A useful fact is that* $\ln(1+x) < x$*, for all* $x \in \mathbb{R}_{>0}$*, easily seen as follows. With* $f(x) = \ln(1+x)$ *and* $g(x) = x$*, note that* f *and* g *are continuous and differentiable, with* $f(0) = g(0) = 0$*, but their slopes are such that* $f'(x) = (1+x)^{-1} < 1 = g'(x)$*, so that, from Example 2.8,* $f(x) < g(x)$ *for all* $x \in \mathbb{R}_{>0}$*.*

We can also prove that $\ln(1+x) < x$ *for* $x > 0$ *(and more) using the MVT. As in Stoll (2021, Example 5.2.7), we wish to prove that*

$$\frac{x}{1+x} \le \ln(1+x) \le x \quad \text{for all} \quad x > -1.$$

Let $f(x) = \ln(1+x)$*,* $x \in (-1,\infty)$*. Then* $f(0) = 0$*. If* $x > 0$*, then by the MVT,* $\exists c \in (0,x)$ *such that*

$$\ln(1+x) = f(x) - f(0) = f'(c)x. \tag{2.99}$$

But $f'(c) = (1+c)^{-1}$ *and* $(1+x)^{-1} < (1+c)^{-1} < 1$ *for all* $c \in (0,x)$*. Therefore*

$$\frac{x}{1+x} < f'(c)x < x, \tag{2.100}$$

and, as a consequence of (2.99) and (2.100), and adding the $x = 0$ *case,*

$$\frac{x}{1+x} \le \ln(1+x) \le x \quad \text{for all} \quad x \ge 0.$$

Now suppose $-1 < x < 0$*. Observe, as* $f(0) = 0$*,*

$$\ln(1+x) = f(x) - f(0) = \frac{-[f(0)-f(x)]}{0-x}(0-x) = x\frac{f(0)-f(x)}{0-x},$$

and, again by the MVT, $\exists c \in (x,0)$ *such that*

$$\frac{f(0)-f(x)}{0-x} = f'(c) = \frac{1}{1+c},$$

i.e., multiplying this by x*,*

$$\ln(1+x) = f(x) - f(0) = \frac{x}{1+c}. \tag{2.101}$$

But as $x < c < 0$*, we have* $1 < (1+c)^{-1} < (1+x)^{-1}$*, and as* x *is negative,*

$$\frac{x}{1+c} > \frac{x}{1+x}.$$

From (2.101) and that $c < 0$*, we have* $1+c < 1$*, so that (as* $x < 0$*)* $\ln(1+x) < x$*. Thus,*

$$\frac{x}{1+x} < \frac{x}{1+c} = \ln(1+x) < x.$$

Hence, the desired inequality holds for all $x > -1$*, with equality if and only if* $x = 0$*.* $\blacksquare$

Example 2.18 *For any $k \in \mathbb{N}$ and $z \in \mathbb{R}$, letting $x = e^z$ shows that*

$$\lim_{x\to\infty} \frac{(\ln x)^k}{x} = \lim_{z\to\infty} \frac{z^k}{e^z} = 0 \tag{2.102}$$

from (2.83). Also, for some $p > 0$, with $z = x^p$, (2.102) implies

$$\lim_{x\to\infty} \frac{\ln x}{x^p} = p^{-1} \lim_{z\to\infty} \frac{(\ln z)}{z} = 0, \tag{2.103}$$

a result required below in Example 2.61. ■

Most students will be familiar with a (common and correct) different definition of the exponential function. Here is the connection. As the derivative of $\ln x$ at $x = 1$ is $1/x = 1$, the Newton quotient (2.32), that $\ln x^p = p \ln x$ from (2.89), and the continuity of the log function combined with continuity result (2.19) imply that

$$1 = \lim_{h\to 0} \frac{\ln(1+h) - \ln 1}{h} = \lim_{h\to 0} \frac{\ln(1+h)}{h} = \lim_{h\to 0} \left(\ln(1+h)^{1/h}\right) = \ln\left(\lim_{h\to 0} (1+h)^{1/h}\right).$$

Taking the inverse function (exponential) and recalling (2.81) gives

$$\exp(1) = e = \lim_{h\to 0} (1+h)^{1/h} = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n.$$

We now prove the other direction. First observe that, from the chain rule and (2.88),

$$\frac{d}{dn} \ln(1 + \lambda/n) = \frac{1}{1+\lambda/n}(-\lambda n^{-2}) = -\frac{\lambda}{n\lambda + n^2}.$$

To evaluate $\lim_{n\to\infty} (1+\lambda/n)^n$, take the log of this, use the continuity of the log function, and l'Hôpital's rule to get

$$\begin{aligned} \ln \lim_{n\to\infty} (1+\lambda/n)^n &= \lim_{n\to\infty} \ln(1+\lambda/n)^n = \lim_{n\to\infty} n \ln(1+\lambda/n) \\ &= \lim_{n\to\infty} \frac{\frac{d}{dn}\ln(1+\lambda/n)}{\frac{d}{dn} n^{-1}} = \lim_{n\to\infty} \frac{-\frac{\lambda}{n\lambda+n^2}}{-\frac{1}{n^2}} = \lambda \lim_{n\to\infty} \left(\frac{n}{n+\lambda}\right) = \lambda, \end{aligned}$$

i.e.,

$$\lim_{n\to\infty} (1+\lambda/n)^n = e^\lambda. \tag{2.104}$$

2.3 Convexity

Here we investigate some basic relations between convexity, continuity, and derivatives. Convexity is an extremely important property in optimization. There are several books dedicated to convexity and optimization. This section is based primarily on Ghorpade and Limaye.

Most students will have learned the following: Given a twice differentiable function $f : \mathbb{R} \to \mathbb{R}$, f is convex if $f''(x) \geq 0$ for all $x \in \mathbb{R}$. Likewise, f is concave if $f''(x) \leq 0$ for all $x \in \mathbb{R}$. Figure 4 shows an illustration of convex and concave functions. For example, $f(x) = ax^2 + bx + c$ is convex if $a \geq 0$, and is concave if $a \leq 0$. These definitions are too specific, requiring f to be twice differentiable.

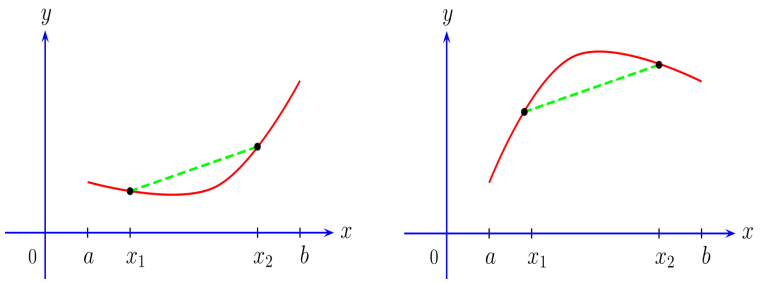


Figure 4: From Ghorpade and Limaye, page 25

We now develop the more general definition, ultimately given below in (2.108), and some basic results.

Geometrically, a function is *convex* if the line segment joining any two points on its graph lies on or above the graph. A function is *concave* if any such line segment lies on or below the graph. Another geometrically visible fact is that, for a convex function, each tangent line of the function lies entirely below the graph of the function. More specifically, Let $f : (a, b) \to \mathbb{R}$ be a convex function. Then, for every point $c \in (a, b)$, one can prove there exists a line L in $\mathbb{R}^2$ with the following properties:

(a) L passes through the point $(c, f(c))$. (b) The graph of f lies entirely above L.

Any line satisfying the above is referred to as a tangent line for f at c. Note that f does not need to be differentiable. If not, then the slope of a tangent line may not be uniquely determined. As an example, consider $f : [0, 1] \to \mathbb{R}$, with $f(x) = x/2$ for $x \in [0, 1/2]$; and $f(x) = x - 1/2$, for $x \in (1/2, 1]$. Another canonical example is $f(x) = |x|$.

Theorem:

$$\text{Every convex function is continuous.} \tag{2.105}$$

> Proof: As in https://e.math.cornell.edu/people/belk/measuretheory/Inequalities.pdf: Let $f : (a, b) \to \mathbb{R}$ be a convex function, and let $c \in (a, b)$. Let L be a linear function whose graph is a tangent line for f at c, and let P be a piecewise linear function consisting of two chords to the graph of f meeting at c. See Figure 5. Then $L \leq f \leq P$ in a neighborhood of c, and $L(c) = f(c) = P(c)$. As L and P are continuous at c, it follows from the Squeeze Theorem and the sequential definition of continuity that f is also continuous at c.

Analytically, for $x_1 < x < x_2$, we think in terms of the slope of the line from x_1 to x, compared to the slope of the line from x_1 to x_2. For convex, the latter should be larger than the former. This gives rise to the following.

Definition: Let $D \subseteq \mathbb{R}$ be such that D contains an interval I, and let $f : D \to \mathbb{R}$ be a function. We say that f is convex on I if

$$x_1, x_2, x \in I, x_1 < x < x_2 \implies f(x) - f(x_1) \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1), \tag{2.106}$$

and f is concave on I if

$$x_1, x_2, x \in I, x_1 < x < x_2 \implies f(x) - f(x_1) \geq \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1). \tag{2.107}$$

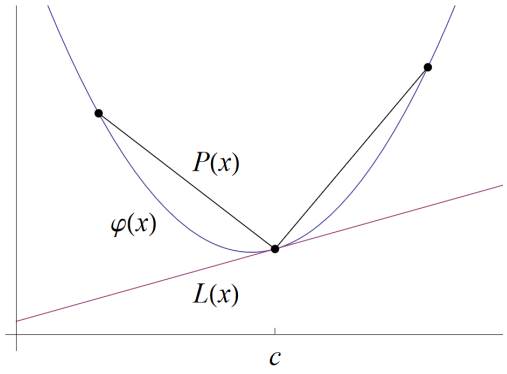


Figure 5: Convex function f is continuous at each point in an open interval of its domain. Taken from https://e.math.cornell.edu/people/belk/measuretheory/Inequalities.pdf.

An alternative way, and the one more commonly seen in the literature, to formulate the definitions of convexity and concavity is as follows.

Proposition: Function f is convex on I if (and only if)

$$\forall x_1, x_2 \in I, \;\; x_1 < x_2, \;\; \forall t \in (0,1), \quad f\left((1-t)x_1 + tx_2\right) \leq (1-t)f\left(x_1\right) + tf\left(x_2\right).$$

Proof: First note that, for all $x_1, x_2 \in \mathbb{R}$ with $x_1 < x_2$, the points x between x_1 and x_2 are of the form $(1-t)x_1 + tx_2$ for some $t \in (0,1)$; in fact, t and x determine each other uniquely, since

$$x = (1-t)x_1 + tx_2 \iff t = \frac{x - x_1}{x_2 - x_1}.$$

Substituting this into the previous definition gives the result.

In the previous result, the roles of t and $1-t$ can be readily reversed, and with this in view, one need not assume that $x_1 < x_2$. Thus, we arrive at our final definition.

Definition: Function f is convex on I if (and only if)

$$f\left(tx_1 + (1-t)x_2\right) \leq tf\left(x_1\right) + (1-t)f\left(x_2\right) \qquad \text{for all } x_1, x_2 \in I \text{ and } t \in (0,1). \tag{2.108}$$

Similarly, f is concave on I if (and only if)

$$f\left(tx_1 + (1-t)x_2\right) \geq tf\left(x_1\right) + (1-t)f\left(x_2\right) \qquad \text{for all } x_1, x_2 \in I \text{ and } t \in (0,1). \tag{2.109}$$

Theorem: A function $f : (a,b) \to \mathbb{R}$ is convex if and only if it is continuous on (a,b) and satisfies

$$f\left(\frac{x_1 + x_2}{2}\right) \leq \frac{f\left(x_1\right) + f\left(x_2\right)}{2}, \quad \forall x_1, x_2 \in (a,b). \tag{2.110}$$

See Ghorpade and Limaye, p. 102, #3.34 for this result. We will in fact prove one direction of (2.110) next, in (2.111), and for a more general linear combination of x_i. The result is well-known, and very important; it is called Jensen's inequality, given in (2.120) below.

Theorem: Let f be convex on (a,b) as in (2.108). Then

$$f \textbf{ is continuous on } (a,b). \tag{2.111}$$

Proof: This is proven without appealing to geometric arguments as above; so purely analytic. The result also holds for f concave. As in Ghorpade and Limaye (Prop 3.15), let I be an open interval in $\mathbb{R}$ and let $f : I \to \mathbb{R}$ be convex on I or concave on I.

First, suppose f is convex. Let $c \in I$. Then there is $r > 0$ such that $[c-r, c+r] \subseteq I$. Let $M := \max\{f(c-r), f(c+r)\}$. For each $x \in [c-r, c+r]$, there is $t \in [0,1]$ such that $x = (1-t)(c-r) + t(c+r)$, and, hence, from (2.108),

$$f(x) \leq (1-t)f(c-r) + tf(c+r) \leq (1-t)M + tM = M. \tag{2.112}$$

Given any $\epsilon > 0$ with $\epsilon \leq 1$, and $x \in \mathbb{R}$, we claim that

$$|x-c| \leq r\epsilon \Longrightarrow x \in I \text{ and } |f(x) - f(c)| \leq \epsilon(M - f(c)).$$

Suppose $|x-c| \leq r\epsilon$. Then $x \in [c-r, c+r]$, since $\epsilon \leq 1$, and so $x \in I$. Define

$$y := c + \frac{x-c}{\epsilon} \quad \text{and} \quad z := c - \frac{x-c}{\epsilon}.$$

Then $|y-c| = |z-c| = |x-c|/\epsilon \leq r$, and so $y, z \in [c-r, c+r]$. Moreover,

$$x = (1-\epsilon)c + \epsilon y \quad \text{and} \quad c = \frac{1}{1+\epsilon}x + \frac{\epsilon}{1+\epsilon}z.$$

Since f is convex and $0 < \epsilon \leq 1$, we see that

$$f(x) \leq (1-\epsilon)f(c) + \epsilon f(y), \quad \text{that is,} \quad f(x) - f(c) \leq \epsilon(f(y) - f(c)). \tag{2.113}$$

Recall $y \in [c-r, c+r]$ and, for each $y \in [c-r, c+r]$, (2.112) implies $f(y) \leq M$. Thus, (2.113) implies that $f(x) - f(c) \leq \epsilon(M - f(c))$. Also, as f is convex and $x, y, z \in [c-r, c+r]$,

$$f(c) \leq \frac{1}{1+\epsilon}f(x) + \frac{\epsilon}{1+\epsilon}f(z), \quad \text{that is,} \quad (1+\epsilon)f(c) \leq f(x) + \epsilon f(z).$$

The last inequality implies that $f(c) - f(x) \leq \epsilon(f(z) - f(c)) \leq \epsilon(M - f(c))$. It follows that $|f(x) - f(c)| \leq \epsilon(M - f(c))$, and thus the claim is established. The result of continuity of f at c follows from the δ-ϵ definition of continuity. If f is concave, it suffices to apply the result just proved to $-f$.

Theorem: Let I be an interval containing more than one point, and let $f : I \to \mathbb{R}$ be a differentiable function. Then

$$\textbf{(i) } f' \textbf{ is monotonically increasing on } I \Longleftrightarrow f \textbf{ is convex on } I. \tag{2.114}$$

Similarly,

(ii) f' is monotonically decreasing on $I \Longleftrightarrow f$ is concave on I.

(iii) f' is strictly increasing on $I \Longleftrightarrow f$ is strictly convex on I.

(iv) f' is strictly decreasing on $I \Longleftrightarrow f$ is strictly concave on I.

Proof of (2.114): As in Ghorpade and Limaye (Prop 4.33). First, assume that f' is monotonically increasing on I. Let $x_1, x_2, x \in I$ be such that $x_1 < x < x_2$. By the MVT, there are $c_1 \in (x_1, x)$ and $c_2 \in (x, x_2)$ satisfying

$$f(x) - f(x_1) = f'(c_1)(x - x_1) \quad \text{and} \quad f(x_2) - f(x) = f'(c_2)(x_2 - x).$$

Now $c_1 < c_2$ and f' is monotonically increasing on I, and so

$$\frac{f(x) - f(x_1)}{x - x_1} = f'(c_1) \leq f'(c_2) = \frac{f(x_2) - f(x)}{x_2 - x}.$$

Collecting only the terms involving $f(x)$ on the left side, we obtain

$$f(x)\left(\frac{1}{x - x_1} + \frac{1}{x_2 - x}\right) \leq \frac{f(x_1)}{x - x_1} + \frac{f(x_2)}{x_2 - x}.$$

Multiplying throughout by $(x - x_1)(x_2 - x)/(x_2 - x_1)$, we see that

$$f(x) \leq \frac{f(x_1)(x_2 - x) + f(x_2)(x - x_1)}{x_2 - x_1} = f(x_1) + \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1),$$

where the last equality follows by writing $x_2 - x = (x_2 - x_1) - (x - x_1)$. Thus, recalling (2.106), f is convex on I.

Conversely, assume that f is convex on I. Let $x_1, x_2, x \in I$ be such that $x_1 < x < x_2$. Then

$$f(x) \leq f(x_1) + \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1) = f(x_1) + \frac{f(x_2) - f(x_1)}{x_2 - x_1}[(x_2 - x_1) - (x_2 - x)]$$
$$= f(x_1) + [f(x_2) - f(x_1)] - \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x_2 - x) = f(x_2) - \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x_2 - x).$$

As a consequence, the slopes of chords are increasing, that is,

$$\frac{f(x) - f(x_1)}{x - x_1} \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq \frac{f(x_2) - f(x)}{x_2 - x}.$$

Taking limits as $x \to x_1^+$ and $x \to x_2^-$, we obtain

$$f'(x_1) \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq f'(x_2)$$

Thus, f' is monotonically increasing on I.

Theorem: Suppose $f''(x)$ exists for all $x \in (a, b)$. Then

$$f \text{ convex on } (a, b) \iff \forall x \in (a, b),\ f''(x) \geq 0. \tag{2.115}$$

Proof: Just apply (2.114) and the slight variant of (2.69), namely: Suppose $f : I \to \mathbb{R}$ is differentiable on the interval I. If $f'(x) \geq 0$ for all $x \in I$, then f is monotone increasing on I.

Theorem: (Ghorpade and Limaye, p. 40, exercise #1.63): Let I be an interval containing more than one point and let $f : I \to \mathbb{R}$ be a function. Define $\phi(x_1, x_2) := (f(x_1) - f(x_2))/(x_1 - x_2)$ for $x_1, x_2 \in I$ with $x_1 \neq x_2$. Then f is convex on I if and only if ϕ is a monotonically increasing function of x_1, that is,

$$\forall x_1, x_2 \in I,\ x_1 < x_2,\ \forall x \in I \setminus \{x_1, x_2\},\quad \phi(x_1, x) \leq \phi(x_2, x). \tag{2.116}$$

Proof: Ghorpade and Limaye do not provide the proof. Here is one.

Montonically increasing if convex: Suppose that f is convex. Assume for contradiction that $\phi(x_1, x)$ is not monotonically increasing in x_1. That is, we can find $x_1, x_2 \in I$ with $x_1 < x_2$ and $x \in I \setminus \{x_1, x_2\}$ such that $\phi(x_1, x) > \phi(x_2, x)$. Assume $x_1 < x < x_2$.[8] Note that, since $x_1 < x < x_2$, $\exists t \in (0,1)$ such that $tx_1 + (1-t)x_2 = x$. Then

$$\begin{aligned}
&\frac{f(x_1) - f(tx_1 + (1-t)x_2)}{x_1 - (tx_1 + (1-t)x_2)} > \frac{f(x_2) - f(tx_1 + (1-t)x_2)}{x_2 - (tx_1 + (1-t)x_2)} \\
\Leftrightarrow &\frac{f(x_1) - f(tx_1 + (1-t)x_2)}{(1-t)(x_1 - x_2)} > \frac{f(x_2) - f(tx_1 + (1-t)x_2)}{-t(x_1 - x_2)} \\
\Leftrightarrow &-t\left[f(x_1) - f(tx_1 + (1-t)x_2)\right] > (1-t)\left[f(x_2) - f(tx_1 + (1-t)x_2)\right] \\
\Leftrightarrow &f(tx_1 + (1-t)x_2) > (1-t)f(x_2) + tf(x_1)
\end{aligned}$$

Note that going from the second to third row we multiply by $-t(1-t)(x_1 - x_2) > 0$, therefore the inequalities do not switch. The last line contradicts that f is convex. We therefore conclude that $\phi(x_1, x)$ is monotonically increasing in x_1.

Convex if montonically increasing: Conversely, suppose that $\phi(x_1, x) \leq \phi(x_2, x)$ for any $x_1, x_2 \in I$ and $x \in I \setminus \{x_1, x_2\}$. Now assume for contradiction that f is not convex, i.e., $\exists t \in (0,1)$ and $x_1, x_2 \in I$ (with $x_1 \neq x_2$)[9] such that $f(tx_1 + (1-t)x_2) > tf(x_1) + (1-t)f(x_2)$. Assume w.l.o.g. that $x_1 < x_2$. Let $x = tx_1 + (1-t)x_2$, then clearly $x \in I$. Then

$$\begin{aligned}
&f(tx_1 + (1-t)x_2) > tf(x_1) + (1-t)f(x_2) \\
\Leftrightarrow\ &f(tx_1 + (1-t)x_2) - f(x_2) > t(f(x_1) - f(x_2)) \\
\Leftrightarrow\ &\frac{f(tx_1 + (1-t)x_2) - f(x_2)}{tx_1 + (1-t)x_2 - x_2} < \frac{t(f(x_1) - f(x_2))}{tx_1 + (1-t)x_2 - x_2} \\
\Leftrightarrow\ &\frac{f(tx_1 + (1-t)x_2) - f(x_2)}{tx_1 + (1-t)x_2 - x_2} < \frac{f(x_1) - f(x_2)}{x_1 - x_2} \\
\Leftrightarrow\ &\phi(x, x_2) < \phi(x_1, x_2)
\end{aligned}$$

The inequality flips in the third line when we divide both sides by $(tx_1 + (1-t)x_2 - x_2) < 0$. The last line contradicts our assumption that $\phi(x_1, x)$ is monotonically increasing in x_1, because $x_1 < x$ but $\phi(x, x_2) < \phi(x_1, x_2)$. Therefore, we conclude that f is convex.

Theorem: If f is convex on (a, b), then

$$f'_+(p) \textbf{ and } f'_-(p) \textbf{ exist for every } p \in (a, b). \tag{2.117}$$

(This appears in Stoll, 2021, p. 221, Misc. Exercise #3, without solution.)

Proof: The case of interest is when f is not differentiable at some $p \in (a, b)$. From (2.105) and (2.111), we know f is continuous at p, i.e., $f(p) = f_+(p) = f_-(p)$. From (2.46) and (2.47), we need to show the existence of

$$f'_+(p) = \lim_{h \to 0^+} \frac{f(p+h) - f(p)}{h} \quad \text{and} \quad f'_-(p) = \lim_{h \to 0^-} \frac{f(p+h) - f(p)}{h}.$$

Recall that a limit of function $f : D \to \mathbb{R}$ as $x \to c$ exists, denoted $f(x) \to \ell$ as $x \to c$, or $\lim_{x \to c} f(x) = \ell$, if there exists $\ell \in \mathbb{R}$ such that, for any sequence $\{x_n\} \in D \setminus \{c\}$ with

$x_n \to c$, $f(x_n) \to \ell$. Consider $f'_+(p)$ and let $x_n = p + 1/n$, for $n \in \{n_0, n_0+1, \ldots\}$, where n_0 is the smallest value of $n \in \mathbb{N}$ such that $p + 1/n < b$ (which we know exists, by invoking the well-ordering principle and the Archimedean Property). Let $h, k \in \mathbb{N}$ such that $n_k > n_h \geq n_0$, so $p < x_{n_k} < x_{n_h}$. From (2.106),

$$\frac{f(x_{n_k}) - f(p)}{x_{n_k} - p} \leq \frac{f(x_{n_h}) - f(p)}{x_{n_h} - p}, \tag{2.118}$$

The result now follows from (2.116). The proof for $f'_-(p)$ is similar, or possibly could be elicited from that of $f'_+(p)$ and some clever "symmetry" argument, defining some function g in terms of f.

As an example that a convex function on (a, b) need not be differentiable on (a, b), consider the following. For any $a > 0$, let $I = [-a, a]$, and $f : I \to \mathbb{R}$ defined by $f(x) = |x|$. Function f is clearly convex, but not differentiable at the interior point $0 \in (-a, a)$. Similarly, $-f$ is concave on I, but not differentiable at 0.

Theorem (Young's inequality): Let $a, b \in \mathbb{R}_{\geq 0}$ and $p, q \in (1, \infty)$ such that $1/p + 1/q = 1$. Then

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}. \tag{2.119}$$

Proof: (Nair, Lemma 5.2.3) Function $\varphi : \mathbb{R} \to \mathbb{R}$ defined by $\varphi(x) = e^x$, $x \in \mathbb{R}$ is convex, i.e., for every $x, y \in \mathbb{R}$ and $0 < \lambda < 1$, $\varphi(\lambda x + (1-\lambda)y) \leq \lambda\varphi(x) + (1-\lambda)\varphi(y)$. Taking $\lambda = 1/p$ we have $1 - \lambda = 1/q$ and

$$e^{x/p + y/q} \leq \frac{e^x}{p} + \frac{e^y}{q}.$$

Now, taking $x > 0$ and $y > 0$ such that $a = e^{x/p}$ and $b = e^{y/q}$, that is, $x = \ln(a^p)$ and $y = \ln(b^q)$, we obtain (2.119).

Theorem (Jensen's inequality, Finite Version): Let $\varphi : (a, b) \to \mathbb{R}$ be a convex function, where $-\infty \leq a < b \leq \infty$, and let $x_1, \ldots, x_n \in (a, b)$. Then

$$\varphi(\lambda_1 x_1 + \cdots + \lambda_n x_n) \leq \lambda_1 \varphi(x_1) + \cdots + \lambda_n \varphi(x_n) \tag{2.120}$$

for any $\lambda_1, \ldots, \lambda_n \in [0, 1]$ satisfying $\lambda_1 + \cdots + \lambda_n = 1$.

NOTE: A more general version that subsumes this case is Jensen's inequality for the Lebesgue integral. An excellent presentation can be found in M. Thamban Nair's *Measure and Integration: A First Course*, Thm 5.2.5. Nair subsequently also shows that Young's inequality is a special case of Jensen.

Proof: As in https://e.math.cornell.edu/people/belk/measuretheory/Inequalities.pdf Let $c = \lambda_1 x_1 + \cdots + \lambda_n x_n$, and let L be a linear function whose graph is a tangent line for φ at c. Since $\lambda_1 + \cdots + \lambda_n = 1$, we know that $L(\lambda_1 x_1 + \cdots + \lambda_n x_n) = \lambda_1 L(x_1) + \cdots + \lambda_n L(x_n)$. As $L \leq \varphi$ and $L(c) = \varphi(c)$, we conclude that

$$\begin{aligned} \varphi(c) = L(c) &= L(\lambda_1 x_1 + \cdots + \lambda_n x_n) \\ &= \lambda_1 L(x_1) + \cdots + \lambda_n L(x_n) \leq \lambda_1 \varphi(x_1) + \cdots + \lambda_n \varphi(x_n). \end{aligned}$$

Theorem (AM-GM Inequality): Let $n \in \mathbb{N}$ and let $a_1, \ldots, a_n$ be nonnegative real numbers. Then the arithmetic mean of $a_1, \ldots, a_n$ is greater than or equal to their geometric mean, that is,

$$\frac{a_1 + \cdots + a_n}{n} \geq \sqrt[n]{a_1 \cdots a_n}. \tag{2.121}$$

Moreover, equality holds if and only if $a_1 = \cdots = a_n$.

Proof: As in Ghorpade and Limaye (Prop 1.11) If some $a_i = 0$, then the result is obvious. Assume $a_i > 0$. Let $g = (a_1 \cdots a_n)^{1/n}$ and $b_i = a_i/g$ for $i = 1, \ldots, n$. Then $b_1, \ldots, b_n$ are positive and $b_1 \cdots b_n = 1$. We shall now show, using induction on n, that $b_1 + \cdots + b_n \geq n$. This is clear if $n = 1$ or if each of $b_1, \ldots, b_n$ equals 1. Suppose $n > 1$ and not every b_i equals 1 . Then $b_1 \cdots b_n = 1$ implies that among $b_1, \ldots, b_n$ there is a number < 1 as well as a number > 1. Relabeling $b_1, \ldots, b_n$ if necessary, we may assume that $b_1 < 1$ and $b_n > 1$. Let $c_1 = b_1 b_n$. Then $c_1 b_2 \cdots b_{n-1} = 1$, and hence by the induction hypothesis $c_1 + b_2 + \cdots + b_{n-1} \geq n - 1$. Now observe that

$$\begin{aligned} b_1 + \cdots + b_n &= (c_1 + b_2 + \cdots + b_{n-1}) + b_1 + b_n - c_1 \\ &\geq (n-1) + b_1 + b_n - b_1 b_n = n + (1 - b_1)(b_n - 1) > n, \end{aligned}$$

because $b_1 < 1$ and $b_n > 1$. This proves that $b_1 + \cdots + b_n \geq n$, and moreover, the inequality is strict unless $b_1 = \cdots = b_n = 1$. Substituting $b_i = a_i/g$, we obtain the desired result.

Theorem (AM-GM Inequality, Unequal Weights): Let $x_1, \ldots, x_n > 0$, and let $\lambda_1, \ldots, \lambda_n \in [0, 1]$ so that $\lambda_1 + \cdots + \lambda_n = 1$. Then

$$x_1^{\lambda_1} \cdots x_n^{\lambda_n} \leq \lambda_1 x_1 + \cdots + \lambda_n x_n. \tag{2.122}$$

Proof: This theorem is equivalent to the convexity of the exponential function. Specifically, from (2.120),

$$\exp\{\lambda_1 t_1 + \cdots \lambda_n t_n\} \leq \lambda_1 e^{t_1} + \cdots + \lambda_n e^{t_n}, \quad \forall t_1, \ldots, t_n \in \mathbb{R}.$$

Substituting $x_i = e^{t_i}$ gives the desired result.

2.4 Integration

> If we evolved a race of Isaac Newtons, that would not be progress. For the price Newton had to pay for being a supreme intellect was that he was incapable of friendship, love, fatherhood, and many other desirable things. As a man he was a failure; as a monster he was superb.
>
> (Aldous Huxley)

Every schoolchild knows the formula for the area of a rectangle. Under certain conditions, the area under a curve can be approximated by summing the areas of adjacent rectangles with heights coinciding with the function under study and ever-decreasing widths. Related concepts go back at least to Archimedes. This idea was of course known to Gottfried Leibniz (1646–1716) and Isaac Newton (1642–1727), though they viewed the integral as an antiderivative (see below) and used it as such. Augustin Louis Cauchy (1789–1857) is credited with using limits of sums as in the modern approach of integration, which led him to prove the fundamental theorem of calculus. Building on the work of Cauchy, Georg Bernhard Riemann (1826–1866) entertained working with discontinuous functions, and ultimately developed the modern definition of what is now called the Riemann integral in 1853, along with necessary and sufficient conditions for its existence. Contributions to its development were also made by Jean Gaston Darboux (1842–1917), while Thomas–Jean Stieltjes (1856–1894) pursued what is now referred to as the Riemann–Stieltjes integral.[10]

2.4.1 Definitions, Existence, and Properties

> The simplest schoolboy is now familiar with facts for which Archimedes would have sacrificed his life. (Ernest Renan)

To make precise the aforementioned notion of summing the area of rectangles, some notation is required. Let $A = [a, b]$, $a < b$, be a bounded interval in $\mathbb{R}$. A *partition* of A is a finite set $\pi = \{x_k\}_{k=0}^n$ such that $a = x_0 < x_1 < \cdots < x_n = b$, and its *mesh* (sometimes called the norm, or size), is given by $\mu(\pi) = \max\{x_1 - x_0, x_2 - x_1, \ldots, x_n - x_{n-1}\}$.

Let π_1 and π_2 be partitions of I.

$$\textbf{If } \pi_1 \subset \pi_2\textbf{, then } \pi_2 \textbf{ is a } \textbf{\textit{refinement}} \textbf{ of } \pi_1. \tag{2.123}$$

A *selection* associated to a partition $\pi = \{x_k\}_{k=0}^n$ is any set $\{\xi_k\}_{k=1}^n$ such that $x_{k-1} \leq \xi_k \leq x_k$ for $k = 1, \ldots, n$.

---

[10] See Stoll (2001, Ch. 6) and Browder (1996, p. 121) for some historical commentary, and Hawkins (1970) for a detailed account of the development of the Riemann and Lebesgue integrals.
The Riemann integral was fundamentally superseded and generalized by the work of Henri Léon Lebesgue (1875–1941) in 1902, as well as Émile Borel (1871–1956) and Constantin Carathéodory (1873–1950), giving rise to the Lebesgue integral. While it is considerably more complicated than the Riemann integral, it has important properties not shared by the latter, to the extent that the Riemann integral is considered by some to be just a historical relic. Somewhat unexpectedly, in the 1950's, Ralph Henstock and Jaroslav Kurzweil independently proposed an integral formulation that generalizes the Riemann integral, but in a more direct and much simpler fashion, without the need for notions of measurable sets and functions, or σ-algebras. It is usually referred to as the gauge integral, or some combination of the pioneers names. It not only nests the Lebesgue integral, but also the improper Riemann and Riemann–Stieltjes integrals. There are several textbooks that discuss it, such as those by or with Charles W. Swartz, and by or with Robert G. Bartle. See the web page by Eric Schechter, `http://www.math.vanderbilt.edu/~schectex/ccc/gauge/` and the references therein for more information.

Now let $f : D \to \mathbb{R}$ with $A \subset D \subset \mathbb{R}$, $\pi = \{x_k\}_{k=0}^n$ be a partition of A, and $\sigma = \{\xi_k\}_{k=1}^n$ a selection associated to π. The *Riemann sum* for function f, with partition π and selection σ, is given by

$$S(f, \pi, \sigma) = \sum_{k=1}^{n} f(\xi_k)(x_k - x_{k-1}). \tag{2.124}$$

Observe how S is just a sum of areas of rectangles with heights dictated by f, π and σ. If the Riemann sum converges to a real number as the level of refinement increases, then f is *integrable*.

Definition: Function f is said to be *(Riemann) integrable* over $A = [a, b]$ if there is a number $I \in \mathbb{R}$ such that: $\forall \epsilon > 0$, there exists a partition π_0 of A such that, for every refinement π of π_0, and every selection σ associated to π, we have $|S(f, \pi, \sigma) - I| < \epsilon$. If f is Riemann integrable over $[a, b]$, then we write $f \in \mathcal{R}[a, b]$.

The number I is called the *integral* of f over $[a, b]$, and denoted by $\int_a^b f$ or $\int_a^b f(x)\,dx$. Observe how, in the latter notation, x is a "dummy variable", in that it could be replaced by any other letter (besides, of course, f, a or b), and also how it mirrors the notation in (2.124), i.e., the term $\sum_{k=1}^n$ is replaced by $\int_a^b$, the term $f(\xi_k)$ is replaced by $f(x)$ and the difference $(x_k - x_{k-1})$ by dx. Indeed, the integral symbol $\int$ is an elongated letter S, for summation, introduced by Leibniz, and the word integral in this context was first used by Jakob Bernoulli.

For f to be integrable, it is necessary (but not sufficient) that f be bounded on $A = [a, b]$. To see this, observe that, if f were *not* bounded on A, then, for every given partition $\pi = \{x_k\}_{k=0}^n$, $\exists k \in \{1, \ldots, n\}$ and an $x \in [x_{k-1}, x_k]$ such that $|f(x)|$ is arbitrarily large. Thus, by varying the element ξ_k of the selection σ associated to π, the Riemann sum $S(f, \pi, \sigma)$ can be made arbitrarily large, and there can be no value I such that $|S(f, \pi, \sigma) - I| < \epsilon$.

Example 2.19 *Let $f(x) = x$. Then the graph of f from 0 to $b > 0$ forms a triangle with area $b^2/2$. For the equally–spaced partition $\pi_n = \{x_k\}_{k=0}^n$, $n \in \mathbb{N}$, with $x_k = kb/n$ and selection $\sigma = \{\xi_k\}_{k=1}^n$ with $\xi_k = x_k = kb/n$, the Riemann sum is*

$$S(f, \pi, \sigma) = \sum_{k=1}^{n} f(\xi_k)(x_k - x_{k-1}) = \sum_{k=1}^{n} \frac{kb}{n}\left(\frac{kb}{n} - \frac{(k-1)\,b}{n}\right),$$

which simplifies to

$$S(f, \pi, \sigma) = \left(\frac{b}{n}\right)^2 \sum_{k=1}^{n} k = \left(\frac{b}{n}\right)^2 \left(\frac{n(n+1)}{2}\right) = \frac{b^2}{2}\frac{(n+1)}{n}. \tag{2.125}$$

This overestimates the area of the triangle because f is increasing and we took the selection $\xi_k = x_k$; likewise, choosing $\xi_k = x_{k-1}$ would underestimate it with $S = \frac{b^2}{2}\frac{(n-1)}{n}$; and, because of the linearity of f, choosing the midpoint $\xi_k = (x_{k-1} + x_k)/2$ gives exactly $b^2/2$. From the boundedness of f on $[a, b]$, the choice of selection will have vanishing significance as n grows, so that, from (2.125), as $n \to \infty$, $S(f, \pi, \sigma) \to b^2/2 = I$. (Of course, to strictly abide by the definition, the partitions would have to be chosen as successive refinements, which is clearly possible.)[11] ■

[11] The more general case with $f(x) = x^p$ for $x \geq 0$ and $p \neq -1$ is particularly straightforward when using a wise choice of non-equally-spaced partition, as was first shown by Pierre De Fermat before the fundamental theorem of calculus was known to him; see Browder (1996, pp. 102, 121) and Stahl (1999, p. 16) for details.

Let $\pi = \{x_k\}_{k=0}^n$ be a partition of f. The *upper (Darboux) sum* of f for π is defined as $\overline{S}(f,\pi) = \sup_\sigma \{S(f,\pi,\sigma)\}$, i.e., the supremum of $S(f,\pi,\sigma)$ over all possible σ associated to π. Likewise, the *lower (Darboux) sum* is $\underline{S}(f,\pi) = \inf_\sigma \{S(f,\pi,\sigma)\}$, and $\underline{S}(f,\pi) \leq \overline{S}(f,\pi)$. By defining

$$m_k = \inf\{f(t) : t \in [x_{k-1}, x_k]\} \quad \text{and} \quad M_k = \sup\{f(t) : t \in [x_{k-1}, x_k]\}, \tag{2.126}$$

we can write $\underline{S}(f,\pi) = \sum_{k=1}^n m_k (x_k - x_{k-1})$ and $\overline{S}(f,\pi) = \sum_{k=1}^n M_k (x_k - x_{k-1})$. Also, if $m \leq f(x) \leq M$ for $x \in [a,b]$, then $m(b-a) \leq \underline{S}(f,\pi) \leq \overline{S}(f,\pi) \leq M(b-a)$. It should be intuitively clear that, if π and π' are partitions of $[a,b]$ such that $\pi \subset \pi'$, then

$$\underline{S}(f,\pi) \leq \underline{S}(f,\pi') \leq \overline{S}(f,\pi') \leq \overline{S}(f,\pi). \tag{2.127}$$

Also, for *any* two partitions π_1 and π_2 of $[a,b]$, let $\pi_3 = \pi_1 \cup \pi_2$ be their *common refinement*, so that, from (2.127), $\underline{S}(f,\pi_1) \leq \underline{S}(f,\pi_3) \leq \overline{S}(f,\pi_3) \leq \overline{S}(f,\pi_2)$, i.e., the lower sum of any partition is less than or equal to the upper sum of any (other) partition. This fact is useful for proving the intuitively plausible result, due to Riemann, but going back to Archimedes (see the Wikipedia entry Method of Exhaustion), and thus sometimes referred to as the Archimedes-Riemann Theorem:

Theorem: If f is a bounded function on $[a,b]$, then

$$\int_a^b f \textbf{ exists iff } \forall \epsilon > 0, \exists\, \pi \textbf{ of } [a,b] \textbf{ s.t. } \overline{S}(f,\pi) - \underline{S}(f,\pi) < \epsilon. \tag{2.128}$$

Proofs can be found in most real analysis books, e.g., Stoll and Fitzpatrick. This, in turn, is used for proving the following important results.

Theorem: If f is a bounded function on $[a,b]$, then

$$\textbf{If } f \textbf{ is monotone on } [a,b], \textbf{ then } \int_a^b f \textbf{ exists}. \tag{2.129}$$

Proof: Let f be a (bounded and) monotone increasing function on $[a,b]$. Let $\pi = \{x_k\}_{k=0}^n$ be a partition of $[a,b]$ with $x_k = a + (k/n)(b-a)$. Then (2.126) implies that $m_k = f(x_{k-1})$ and $M_k = f(x_k)$, and, as $x_k - x_{k-1} = (b-a)/n$,

$$\begin{aligned} \overline{S}(f,\pi) - \underline{S}(f,\pi) &= \sum_{k=1}^n [f(x_k) - f(x_{k-1})](x_k - x_{k-1}) \\ &= \frac{b-a}{n} \sum_{k=1}^n [f(x_k) - f(x_{k-1})] = \frac{b-a}{n}(f(b) - f(a)). \end{aligned}$$

As f is bounded and increasing, $0 \leq f(b) - f(a) < \infty$, and n can be chosen such that the rhs is less than any $\epsilon > 0$. Thus, by (2.128), $\int_a^b f$ exists.

Theorem: A continuous function on a closed and bounded interval is integrable:

$$\textbf{If } f \in \mathcal{C}^0[a,b], \textbf{ then } \int_a^b f \textbf{ exists}. \tag{2.130}$$

Proof: Recall from (2.27) that continuity of f on a closed, bounded interval I implies that f is uniformly continuous on I. Thus, $\forall \epsilon > 0$, $\exists \delta > 0$ such that $|f(x) - f(y)| < \epsilon/(b-a)$ when $|x-y| < \delta$. Let $\pi = \{x_k\}_{k=0}^n$ be a partition of $[a,b]$ with mesh $\mu(\pi) < \delta$. Then, for any values $s, t \in [x_{k-1},\ x_k]$, $|s-t| < \delta$ and $|f(s) - f(t)| < \epsilon/(b-a)$. In particular, from (2.126), $M_k - m_k \leq \epsilon/(b-a)$ (the strict inequality is replaced with $\leq$ because of the nature of inf and sup) and

$$\overline{S}(f,\pi) - \underline{S}(f,\pi) = \sum_{k=1}^{n} (M_k - m_k)(x_k - x_{k-1}) \leq \frac{\epsilon}{b-a} \sum_{k=1}^{n} (x_k - x_{k-1}) = \epsilon.$$

Thus, by (2.128), $\int_a^b f$ exists.

Theorem (Riemann-Lebesgue): If f is a bounded function on $[a,b]$ whose set of discontinuities has measure zero, then

$$\int_a^b f \text{ exists.} \tag{2.131}$$

See Browder (1996, p. 104) for a short, easy proof when there exists a measure zero cover C for the set of discontinuity points, and C consists of a *finite* set of disjoint open intervals. This restriction to a finite set of intervals can be lifted, and is referred to as *Lebesgue's Theorem*, given by him in 1902. See e.g., Stoll (2001, §6.7) or Pugh (2002, pp. 165-7) for detailed proofs, and Terrell, §12.5 for the proof for the multivariate Riemann integral, the theorem of which we state in §5.2.3.

Theorem: Let f and ϕ be functions such that $f \in \mathcal{R}[a,b]$, with $m \leq f(x) \leq M$ for all $x \in [a,b]$, and $\phi \in \mathcal{C}^0[m,M]$. Then

$$\phi \circ f \in \mathcal{R}[a,b]. \tag{2.132}$$

See, e.g., Browder (1996, pp. 106-7) or Stoll (2001, pp. 217-8) for elementary proofs, and Stoll (2001, p. 220) or Pugh (2002, p. 168) for the extremely short proof using Lebesgue's theorem.

Valuable special cases include $\phi(y) = |y|$ and $\phi(y) = y^2$, i.e., if f is integrable on $I = [a,b]$, then so are $|f|$ and f^2 on I. It is *not* necessarily true that the composition of two integrable functions is integrable.

We now state some important properties. (Proofs can be found in all real analysis textbooks.) With $f, g \in \mathcal{R}[a,b]$ and $I = [a,b]$, we have *monotonicity*, i.e.,

$$\text{if } f(x) \leq g(x) \text{ for all } x \in I, \text{ then } \int_a^b f(x)\,dx \leq \int_a^b g(x)\,dx. \tag{2.133}$$

For any constants $k_1, k_2 \in \mathbb{R}$, we have *linearity*, or *additivity* and *homogeneity*:

$$\int_a^b (k_1 f + k_2 g) = k_1 \int_a^b f + k_2 \int_a^b g, \tag{2.134}$$

with obvious extension to the sum of $n \in \mathbb{N}$ such integrals. These can be used to prove (the also intuitively obvious)

$$\left| \int_a^b f \right| \leq \int_a^b |f|. \tag{2.135}$$

If $f, g \in \mathcal{R}[a,b]$, then

$$fg \in \mathcal{R}[a,b] \tag{2.136}$$

(see, e.g., Ghorpade and Limaye, Prop 6.16 (iii) for proof), and

$$\left(\int_a^b fg\right)^2 \leq \left(\int_a^b f^2\right)\left(\int_a^b g^2\right), \tag{2.137}$$

known as the *Schwarz*, *Cauchy–Schwarz*, or *Bunyakovsky–Schwarz* inequality.[12]

Let $a < c < b$ and let f be a function on $[a,b]$. Then, $\int_a^b f$ exists iff $\int_a^c f$ and $\int_c^b f$ exist, in which case

$$\int_a^b f = \int_a^c f + \int_c^b f, \tag{2.138}$$

which is referred to as *domain additivity*, or *additivity over partitions*. Also, as definitions,

$$\int_a^a f := 0 \quad \text{and} \quad \int_b^a := -\int_a^b f. \tag{2.139}$$

The motivation for the former is obvious; for the latter definition, one reason is so that (2.138) holds even for $c < a < b$. That is,

$$\int_a^b f = \int_a^c f + \int_c^b f = -\int_c^a f + \int_c^b f = \int_c^b f - \int_c^a f,$$

which corresponds to our intuitive notion of working with pieces of areas.

The *Mean Value Theorem for Integrals* states: Let $f, p \in \mathcal{C}^0(I)$ for $I = [a,b]$, with p nonnegative. Then $\exists c \in I$ such that

$$\int_a^b f(x)p(x)\,dx = f(c)\int_a^b p(x)\,dx. \tag{2.140}$$

A popular and useful form of the theorem just takes $p(x) \equiv 1$, so that $\int_a^b f = f(c)(b-a)$. To prove (2.140), use (2.29) to let

$$m = \min\{f(t) : t \in I\} \quad \text{and} \quad M = \max\{f(t) : t \in I\}. \tag{2.141}$$

As $p(t) \geq 0$, $mp(t) \leq f(t)p(t) \leq Mp(t)$ for $t \in I$, so that

$$m\int_a^b p(t)\,dt \leq \int_a^b f(t)p(t)\,dt \leq M\int_a^b p(t)\,dt,$$

or, assuming $\int_a^b p(t)\,dt > 0$, $m \leq \gamma \leq M$ where $\gamma = \int_a^b f(t)p(t)\,dt \,/\, \int_a^b p(t)\,dt$. From (2.141) and the IVT (2.30), $\exists c \in I$ such that $f(c) = \gamma$, implying (2.140).

The *Bonnet Mean Value Theorem* states that, if $f, g \in \mathcal{C}^0(I)$ and f is positive and decreasing, then $\exists\, c \in I$ such that

$$\int_a^b f(x)g(x)\,dx = f(a)\int_a^c g(x)\,dx. \tag{2.142}$$

It is credited to Pierre Ossian Bonnet (1819–1892), who discovered it in 1849. Lang (1997, p. 107) provides an outline of the proof. There is a similar statement to (2.142) for f positive and increasing. Oddly, most real analysis books do not include this result. A related result, termed the first mean value theorem for integrals, is given in Ghorpade and Limaye (2018, p. 231).

[12] Bunyakovsky was first, having published the result in 1859, while Schwarz found it in 1885 (see Browder, 1996, p. 121). The finite-sum analog of (2.137) is (1.9).

2.4.2 Fundamental Theorem of Calculus

Definition: Let $f : I \to \mathbb{R}$. The function $F : I \to \mathbb{R}$ is called an *antiderivative*, or a *primitive* of f if, $\forall x \in I$, $F'(x) = f(x)$.

The fundamental theorem of calculus, or, in short, FTC, is the link between the differential and integral calculus, of which there are two forms, say FTC (i) and FTC (ii).

Theorem (FTC i): For $f \in \mathcal{R}[a,b]$ with F a primitive of f,

$$\int_a^b f = F(b) - F(a). \tag{2.143}$$

Proof: Let $\pi = \{x_k\}_{k=0}^n$ be a partition of $I = [a,b]$. From the definition, $F'(t) = f(t)$ for all $t \in I$. Applying the MVT to F implies that $\exists \xi_k \in (x_{k-1},\ x_k)$ such that

$$F(x_k) - F(x_{k-1}) = F'(\xi_k)(x_k - x_{k-1}) = f(\xi_k)(x_k - x_{k-1}).$$

The set of ξ_k, $k = 1, \ldots n$, forms a selection, $\sigma = \{\xi_k\}_{k=1}^n$, associated to π, so that

$$S(f,\pi,\sigma) = \sum_{k=1}^n f(\xi_k)(x_k - x_{k-1}) = \sum_{k=1}^n (F(x_k) - F(x_{k-1})) = F(b) - F(a).$$

This holds for any partition π, so that $\int_a^b f(t)\,dt = F(b) - F(a)$.

Observe from (2.143), (2.41), and that, $\forall x \in I$, $F'(x) = f(x)$, that

$$\frac{d}{db}\int_a^b f = \frac{d}{db}F(b) - \frac{d}{db}F(a) = f(b). \tag{2.144}$$

Example 2.20 *Let $I = [a,b]$, $f : I \to \mathbb{R}$ be a continuous (and thus integrable) function on I, and $x \in I$. Differentiating the relation $\int_a^b f = \int_a^x f + \int_x^b f$ w.r.t. x and using (2.144) gives*

$$0 = \frac{d}{dx}\left(\int_a^x f\right) + \frac{d}{dx}\left(\int_x^b f\right) = f(x) + \frac{d}{dx}\left(\int_x^b f\right),$$

which implies that

$$\frac{d}{dx}\left(\int_x^b f(t)\,dt\right) = -f(x), \quad x \in I. \qquad \blacksquare$$

Theorem (FTC ii, a, b): For $f \in \mathcal{R}[a,b]$, define $F(x) = \int_a^x f$, $x \in I = [a,b]$. Then (a):

$$F \in \mathcal{C}^0[a,b], \tag{2.145}$$

and, if f is continuous at $x \in I$, then (b):

$$F'(x) = f(x). \tag{2.146}$$

Proof: For part (a), i.e., (2.145), we demonstrate the $\epsilon - \delta$ formulation of continuity, as given in (2.26). That is, for any given $\epsilon > 0$, and any given $x \in I$, $\exists \delta > 0$ such that

$$y \in I \text{ and } |y - x| < \delta \Longrightarrow |F(y) - F(x)| < \epsilon. \tag{2.147}$$

Fix an $\epsilon > 0$ and $x \in I$, and let $M = \sup\{|f(t)| : t \in I\} \geq 0$. Then, for any $h \in \mathbb{R}$ with $y = x + h \in I$ (in particular, h can be negative, so $x = b$ is allowed), $|y - x| = |h|$, and

$$\begin{aligned} |F(x+h) - F(x)| &= \left| \int_a^{x+h} f(t)\, dt - \int_a^x f(t)\, dt \right| \\ &= \left| \int_x^{x+h} f(t)\, dt \right| \leq \int_x^{x+h} |f(t)|\, dt \leq M|h|. \end{aligned} \tag{2.148}$$

By choosing $\delta = \epsilon/M$ and taking y such that $|h| = |y - x| < \delta$, (2.148) becomes

$$|F(x+h) - F(x)| \leq M|h| = M|y - x| < M\delta = M(\epsilon/M) = \epsilon,$$

which is (2.147). Thus, F is continuous at x. As x was chosen to be any value in I, F is continuous on I.

To prove (2.146), note that, from (2.26), if f is continuous at $x \in I$, then, for any $\epsilon > 0$, $\exists \delta > 0$ such that, for $t \in I$ with $|t - x| < \delta$, $|f(t) - f(x)| < \epsilon$.

First consider the case of establishing the result for $x \in I \setminus \{b\} = [a, b)$. We can always find an $h > 0$ such that $x + h \in [a, b)$. Choose h such that $0 < h < \delta$. Then $\int_x^{x+h} dt = h$, $h^{-1} \int_x^{x+h} f(y)\, dt = f(y)$, and, using inequality (2.135),

$$\begin{aligned} \left| \frac{F(x+h) - F(x)}{h} - f(x) \right| &= \left| \frac{1}{h} \int_x^{x+h} f(t)\, dt - f(x) \right| = \left| \frac{1}{h} \int_x^{x+h} [f(t) - f(x)]\, dt \right| \\ &\leq \frac{1}{h} \int_x^{x+h} |f(t) - f(x)|\, dt < \frac{\epsilon}{h} \int_x^{x+h} dt = \epsilon, \end{aligned}$$

showing that, for $x \in [a, b)$, $F'(x) = f(x)$.

We could consider $x \in I \setminus \{a\} = (a, b]$ and $h < 0$ such that $x + h \in (a, b]$. Instead, we develop the general proof valid for $x \in I$ and $h \in \mathbb{R} \setminus \{0\}$ (negative or positive), and can always be chosen such that $x + h \in I$. Let $|h| < \delta$. Then, noting that $\int_x^{x+h} dt = h$,

$$\begin{aligned} \left| \frac{F(x+h) - F(x)}{h} - f(x) \right| &= \left| \frac{1}{h} \int_x^{x+h} f(t)\, dt - f(x) \right| = \frac{1}{|h|} \left| \int_x^{x+h} [f(t) - f(x)]\, dt \right| \\ &\leq \frac{1}{|h|} \left| \int_x^{x+h} |f(t) - f(x)|\, dt \right| < \frac{\epsilon}{|h|} \left| \int_x^{x+h} dt \right| = \frac{\epsilon}{|h|} |h|. \end{aligned}$$

Thus, $\forall x \in [a, b]$, $F'(x) = f(x)$. It is also easy to show that (2.146) implies (2.143) when f is continuous; see Browder (1996, p. 112) or Priestley (1997, Thm. 5.5).

An informal, graphical proof of (2.146) is of great value for remembering and understanding the result, as well as for convincing others. The left panel of Figure 6 shows a plot of part of a continuous function, $f(x) = x^3$, with vertical lines indicating $x = 3$ and $x = 3.1$. This is "magnified" in the middle panel, with a vertical line at $x = 3.01$, but keeping the same scaling on the y–axis to emphasize the approximate linearity of the function over a relatively

small range of x–values. The rate of change, via the Newton quotient, of the area $A(t)$ under the curve from $x = 1$ to $x = t = 3$ is

$$\frac{A(t+h) - A(t)}{h} = \frac{\int_1^{3+h} f - \int_1^3 f}{h} = \frac{1}{h}\int_3^{3+h} f \approx \frac{h\, f(t)}{h},$$

because, as $h \to 0$ the region under study approaches a rectangle with base h and height $f(t)$; see the right panel of Figure 6.

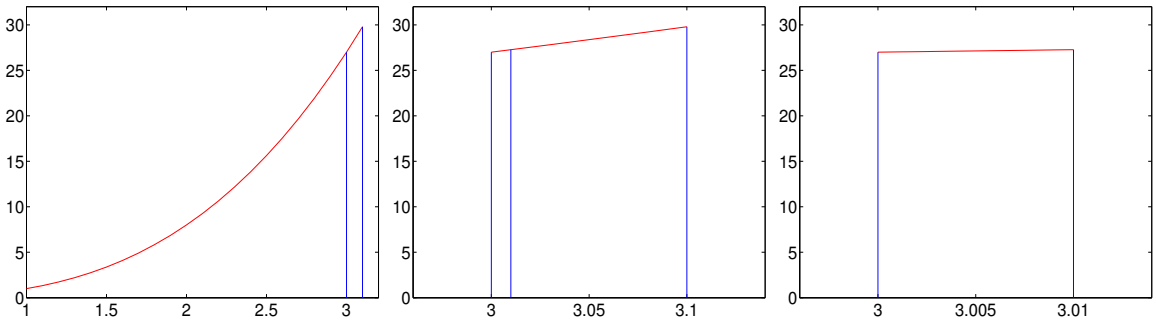


Figure 6: Graphical illustration of the FTC (2.146).

Theorem:

Any two primitives of f differ by a constant. (2.149)

Proof: As in Hijab (1997, p. 103), let F and G be primitives of f on (a, b), so that $H = F - G$ is a primitive of zero, i.e., $\forall x \in (a, b)$, $H'(x) = (F(x) - G(x))' = 0$. The MVT (2.63) implies that $\exists c \in (a, b)$ such that, for $a < x < y < b$,

$$H(x) - H(y) = H'(c)(x - y) = 0,$$

i.e., H is a constant. Thus, any two primitives of f differ by a constant.

Before proceeding to the next example, we need the concept of an indefinite integral. First recall FTC (i), i.e., for $f \in \mathcal{R}[a, b]$ with F a primitive of f, $\int_a^b f = F(b) - F(a)$.

Definition: If an integrable function $f : [a, b] \to \mathbb{R}$ has a primitive F, then F is called an *indefinite integral of* f, and it is denoted by $\int f(x)dx$. This notation is ambiguous, because, from (2.149), a primitive of f is unique only up to an additive constant. For this reason, one writes

$$\int f(x)\, dx = F(x) + C, \tag{2.150}$$

where C denotes an arbitrary constant. Notice that, in this case, $\int_a^b f(x)dx = F(b) - F(a)$, where the right side is independent of the choice of an indefinite integral. The right side of the above equality is sometimes denoted by $[F(x)]_a^b$ or $F(x)|_a^b$. With this in mind, the Riemann integral of $f : [a, b] \to \mathbb{R}$ is sometimes referred to as the *definite integral of* f *over* $[a, b]$. As a simple example, for $f(x) = x$, an antiderivative of f is $F(x) = x^2/2$, so (2.150) implies $\int f(x)dx = F(x) + C$, where C is an arbitrary number. We have $\int_0^1 f = F(x)|_a^b = 1^2/2 + C - (0^2/2 + C) = 1/2$.

Example 2.21 *(Bóna and Shabanov, Concepts in Calculus II, 2012, p. 32) To compute* $\int \cos^4 x\,dx$*, use* (2.54) *to get that* $\cos^2 x = (1+\cos 2x)/2$*, and so*

$$\cos^4 x = \left(\frac{1+\cos 2x}{2}\right)^2 = \frac{1}{4} + \frac{\cos 2x}{2} + \frac{\cos^2 2x}{4}.$$

Applying (2.54) *again, with* $2x$ *replacing* x*, we get that* $\cos^2 2x = (1+\cos 4x)/2$*, so*

$$\cos^4 x = \frac{3}{8} + \frac{\cos 2x}{2} + \frac{\cos 4x}{8}.$$

From the FTC (i) (2.143) *but via* (2.150)*; and* (2.55)*,*

$$\int \cos^4 x\,dx = \int \left(\frac{3}{8} + \frac{\cos 2x}{2} + \frac{\cos 4x}{8}\right) dx = \frac{3x}{8} + \frac{\sin 2x}{4} + \frac{\sin 4x}{32} + C. \qquad \blacksquare$$

A vastly useful technique for resolving integrals is the *change of variables*.

Theorem (Integration by Substitution): Let $I = [a,b]$ for $a < b$, and let $f : I \to \mathbb{R}$ be continuous. Let $\phi : [\alpha,\beta] \to \mathbb{R}$ be such that (i) $\phi([\alpha,\beta]) = [a,b] = I$; (ii) ϕ is differentiable; and (iii) ϕ' is integrable on $[\alpha,\beta]$. Then $(f\circ\phi)\phi' : [\alpha,\beta] \to \mathbb{R}$ is integrable and

$$\int_{\phi(\alpha)}^{\phi(\beta)} f(x)dx = \int_\alpha^\beta f(\phi(t))\phi'(t)dt. \tag{2.151}$$

Proof: From (2.130), f is integrable. Define $F : I \to \mathbb{R}$ by $F(x) = \int_a^x f(u)du$. From FTC (ii, b) in (2.146), F is differentiable, and, $\forall x \in I$, $F'(x) = f(x)$. Next define $H : [\alpha,\beta] \to \mathbb{R}$ by $H = F\circ\phi$. As F and ϕ are differentiable, the chain rule (2.40) implies

$$\forall t \in [\alpha,\beta],\quad H'(t) = F'(\phi(t))\phi'(t) = f(\phi(t))\phi'(t),\quad \text{i.e.,}\quad H' = (f\circ\phi)\phi'.$$

Since ϕ is differentiable, it is continuous, and since f is also continuous, from (2.25), the composite $f\circ\phi$ is continuous and hence, from (2.130), integrable. As ϕ' is integrable, (2.136) implies that $H' = (f\circ\phi)\phi'$ is integrable. Hence, from FTC (i) in (2.143),

$$\int_\alpha^\beta H'(t)dt = H(\beta) - H(\alpha) = \int_a^{\phi(\beta)} f(x)dx - \int_a^{\phi(\alpha)} f(x)dx = \int_{\phi(\alpha)}^{\phi(\beta)} f(x)dx,$$

where the last equality follows from domain additivity (2.138). This proves (2.151).

Example 2.22 *(Sasane, p. 224)*[13] *Consider the integral* $\int_0^1 t\sqrt{1-t^2}\,dt$*. Let* $u = \phi(t) = 1-t^2$*,* $t \in [0,1]$*, and take* $f(u) = \sqrt{u}$*,* $u \in [0,1]$*, so that* $du = -2tdt$*,* $tdt = -du/2$*,* $t = 0 \Rightarrow u = 1$*, and* $t = 1 \Rightarrow u = 0$*. Thus, from* (2.151)*,*

$$\begin{aligned}\int_0^1 t\sqrt{1-t^2}\,dt &= \int_1^0 \sqrt{u}\left(-\frac{1}{2}\right)du = \frac{1}{2}\int_0^1 \sqrt{u}\,du \\ &= \frac{1}{2}\cdot\frac{1}{1+\frac{1}{2}}u^{1+\frac{1}{2}}\Big|_0^1 = \frac{1}{2}\cdot\frac{2}{3}\cdot\left(1^{3/2}-0^{3/2}\right) = \frac{1}{3}. \qquad \blacksquare\end{aligned}$$

[13] The How and Why of One Variable Calculus, 2015

Example 2.23 *(Sasane, p. 224) Consider the integral* $\int_0^{\pi/2}(\sin t)^5\cos t\,dt$. *Let* $u=\phi(t)=\sin t$, $t\in\left[0,\frac{\pi}{2}\right]$, *and* $f(u)=u^5$, $u\in[0,1]$, $du=\cos t dt$, $t=0\Rightarrow u=0$, $t=\frac{\pi}{2}\Rightarrow u=1$. *Thus*

$$\int_0^{\pi/2}(\sin t)^5\cos t\,dt=\int_0^1 u^5 du=\frac{1}{6}u^6\Big|_0^1=\frac{1}{6}. \qquad \blacksquare$$

Example 2.24 *(Sasane, p. 225) Consider the integral* $\int_2^5\frac{1}{t\log t}\,dt$. *Let* $u=\phi(t)=\log t$, $t\in[2,5]$, *and* $f(u)=1/u$ *for* $u\in[\log 2,\log 5]$, $du=dt/t$, $t=2\Rightarrow u=\log 2$, $t=5\Rightarrow u=\log 5$. *Thus, from* (2.88) *and FTC (i) in* (2.143),

$$\int_2^5\frac{1}{t\log t}\,dt=\int_{\log 2}^{\log 5}\frac{1}{u}du=\log u|_{\log 2}^{\log 5}=\log(\log 5)-\log(\log 2). \qquad \blacksquare$$

Example 2.25 *(Bóna and Shabanov, p. 33) To compute* $\int\sin^3 x\,dx$, *write*

$$\sin^3 x=\sin x\cdot\sin^2 x=\sin x\cdot\left(1-\cos^2 x\right)=\sin x-\sin x\cos^2 x,$$

so that, with $u=\cos x$, $du/dx=-\sin x$, *and*

$$\int-\sin x\cos^2 x\,dx=\int u^2 du=\frac{u^3}{3}+C=\frac{\cos^3 x}{3}+C.$$

As $\int\sin x\,dx=-\cos x$, *we get*

$$\int\sin^3 x\,dx=-\cos x+\frac{\cos^3 x}{3}+C. \qquad \blacksquare$$

Many more examples similar to Examples 2.24 and 2.25 can be found at https://math.libretexts.org/Bookshelves/Calculus/Calculus_(OpenStax)/07%3A_Techniques_of_Integration/7.02%3A_Trigonometric_Integrals

Example 2.26 *(Bóna and Shabanov, p. 37) For the integral* $\int(1+x^2)^{-2}\,dx$, *substitute* $x=\tan y$, *so* $y=\tan^{-1}(x)$, *and from* (2.61), $dy=dx/(1+x^2)$. *This yields*

$$\begin{aligned}\int\frac{dx}{(1+x^2)^2}&=\int\frac{dy}{1+x^2}=\int\frac{dy}{1+\tan^2 y}\quad(\tan=\sin/\cos)\\&=\int\cos^2 y\,dy=\frac{y}{2}+\frac{\sin 2y}{4}=\frac{y}{2}+\frac{\sin y\cos y}{2}+C,\end{aligned}$$

where $\int\cos^2 y\,dy$ *is resolved as shown in Example 2.21, and then having used* (2.48a). *Now noting that*

$$\frac{x}{x^2+1}=\frac{\tan y}{1+\tan^2 y}=\tan y\cos^2 y=\sin y\cos y,$$

we have

$$\int f:=\int\frac{dx}{(1+x^2)^2}=\frac{y}{2}+\frac{\sin y\cos y}{2}=\frac{1}{2}\cdot\tan^{-1}(x)+\frac{1}{2}\cdot\frac{x}{x^2+1}+C.$$

Indeed, differentiating the rhs (call it F) gives

$$\begin{aligned} F' &= \frac{1}{2}\frac{1}{1+x^2} + \frac{1}{2}\left(\frac{(x^2+1)-x\,(2x)}{(1+x^2)^2}\right) = \frac{1}{2x^2+2} + \frac{1-x^2}{4x^2+2x^4+2} \\ &= \left(\frac{1}{2x^2+2}\right)\left(\frac{4x^2+2x^4+2}{4x^2+2x^4+2}\right) + \left(\frac{1-x^2}{4x^2+2x^4+2}\right)\left(\frac{2x^2+2}{2x^2+2}\right) \\ &= \frac{4x^2+2x^4+2}{(2x^2+2)\,(4x^2+2x^4+2)} + \frac{2-2x^4}{(2x^2+2)\,(4x^2+2x^4+2)} \\ &= \frac{1}{(2x^2+x^4+1)} = \frac{1}{(1+x^2)^2} = f. \qquad \blacksquare \end{aligned}$$

Example 2.27 *(Sasane, p. 280) Consider the circle given by $x^2+y^2=r^2$, where $r>0$. The area of the circular disk enclosed by the circle is the area of the region between the graphs of the functions $f^+(x) := \sqrt{r^2-x^2}$ and $f_-(x) := -\sqrt{r^2-x^2}$. Thus the area of the disk is*

$$\text{Area}(R) = \int_{-r}^{r}\left(\sqrt{r^2-x^2} - \left(-\sqrt{r^2-x^2}\right)\right)dx = 2\int_{-r}^{r}\sqrt{r^2-x^2}\,dx = 4\int_0^r \sqrt{r^2-x^2}\,dx,$$

where the last equality follows because $x \mapsto \sqrt{r^2-x^2}$ is an even function. We now use the substitution $x = r\cos\theta$, so that $dx = -r\sin\theta d\theta$, and when $x=0$, we have $\theta = \pi/2$, while if $x=r$ then we have $\theta = 0$. So we obtain, using (2.53), i.e., $\cos 2x = 1-2\sin^2 x$,

$$\begin{aligned} \text{Area}(R) &= 4\int_0^r \sqrt{r^2-x^2}\,dx = 4\int_{\pi/2}^0 \sqrt{r^2-r^2(\cos\theta)^2}\cdot(-r\sin\theta)d\theta \\ &= 4r^2\int_0^{\pi/2}(\sin\theta)^2 d\theta = 2r^2\int_0^{\pi/2}(1-\cos(2\theta))d\theta \\ &= 2r^2\left(\frac{\pi}{2} - \frac{\sin(2\theta)}{2}\Big|_0^{\pi/2}\right) = 2r^2\left(\frac{\pi}{2}-0\right) = \pi r^2, \end{aligned} \tag{2.152}$$

i.e., πr^2 is the area of a circular disk of radius r. $\blacksquare$

Example 2.28 *(Ghorpade and Limaye, 2018, Prop 8.2) Let a, b be positive real numbers. We wish to show: The area of the region enclosed by an ellipse given by $(x^2/a^2)+(y^2/b^2)=1$ is equal to πab. As an important special case, setting $b=a$, we see that the area of a disk of radius a is equal to πa^2, as in (2.152).*

Proof: The area enclosed by the given ellipse is four times the area between the curves given by $y = b\sqrt{a^2-x^2}/a$, $y=0$ and between the lines given by $x=0$, $x=a$. Hence (with explanations following) it is equal to

$$4\frac{b}{a}\int_0^a \sqrt{a^2-x^2}\,dx = \frac{4b}{a}\cdot a^2\int_0^{\pi/2}\cos^2\theta\,d\theta = 4ab\int_0^{\pi/2}\frac{1+\cos 2\theta}{2}d\theta = \pi ab, \tag{2.153}$$

where the first equality is obtained from substituting $x = a\sin\theta$, $\theta = \arcsin(x/a)$, $dx = a\cos\theta\,d\theta$, so that

$$\int_0^a \sqrt{a^2-x^2}\,dx = \int_0^{\pi/2}\sqrt{a^2-a^2\sin^2\theta}\,a\cos\theta\,d\theta = a^2\int_0^{\pi/2}\cos^2\theta\,d\theta;$$

the second equality is from (2.54)*; and the third follows from FTC (i) in* (2.143)*, and*

$$\int_0^{\pi/2} \cos(2\theta)d\theta = \frac{1}{2}\int_0^{\pi} \cos(u)du = \frac{1}{2}[\sin\pi - \sin 0] = 0,$$

having used substitution $u = 2\theta$ *and* (2.151). ■

For computing the volume of a three-dimensional object, we need the concept of slices. We use the presentation from Ghorpade and Limaye, 2018, p. 302.

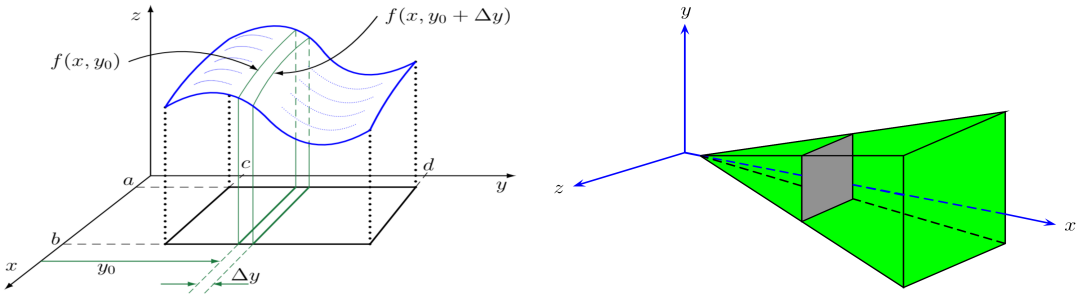


Figure 7: Adding slices to determine volume. Left is from Miklavcic, An Illustrative Guide to Multivariable and Vector Calculus (2020, p. 185); right is from Ghorpade and Limaye, 2018, p. 302.

Let D be a bounded subset of $\mathbb{R}^3 := \{(x, y, z) : x, y, z \in \mathbb{R}\}$ lying between two parallel planes and let L denote a line perpendicular to these planes. A cross-section of D by a plane is called a *slice* of D. See Figure 7. Let us assume that we are able to determine the "area" of a slice of D by any plane perpendicular to L.

For the sake of concreteness, let the line L be the x-axis and assume that D lies between the planes given by $x = a$ and $x = b$, where $a, b \in \mathbb{R}$ with $a < b$. For $s \in [a, b]$, let $A(s)$ denote the area of the slice $\{(x, y, z) \in D : x = s\}$ obtained by intersecting D with the plane given by $x = s$. If $\{x_0, x_1, \ldots, x_n\}$ is a partition of $[a, b]$, then the solid D gets divided into n subsolids

$$\{(x, y, z) \in D : x_{i-1} \le x \le x_i\}, \quad i = 1, \ldots, n.$$

Let us choose $s_i \in [x_{i-1}, x_i]$ and replace the ith subsolid by a rectangular slab having volume equal to $A(s_i)(x_i - x_{i-1})$ for $i = 1, \ldots, n$. Then it is natural to consider

$$\sum_{i=1}^{n} A(s_i)(x_i - x_{i-1})$$

as an approximation of the desired volume of D. We therefore define the volume of D to be

$$\text{Vol}(D) := \int_a^b A(x)dx, \tag{2.154}$$

provided the "area function" $A : [a, b] \to \mathbb{R}$ is integrable.

Example 2.29 *(Bóna and Shabanov, p. 11) Let S be the right circular cone whose symmetry axis is the y axis, whose apex is at $y = h$, and whose base is a circle in the plane $y = 0$ with its center at the origin and with radius r. See Figure 8. Find $V(S)$, the volume of S.*

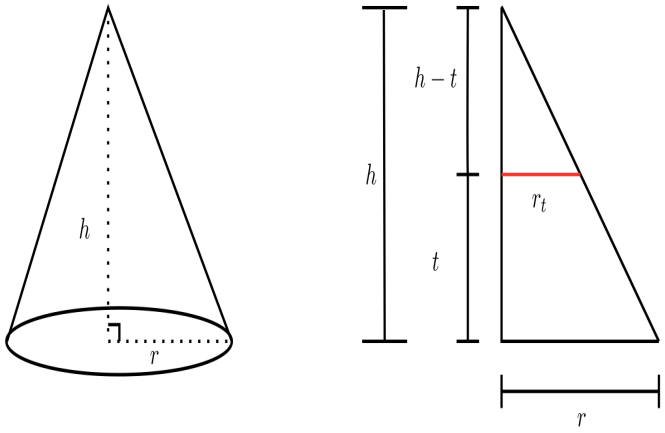


Figure 8: Right circular cone, and similar triangles

Solution: The cone S is between the planes $y = 0$ and $y = h$. The intersection of the plane $y = t$ and S is a circle. The radius r_t of this circle, by similar triangles, satisfies

$$\frac{r_t}{r} = \frac{h-t}{h},$$

showing that $r_t = r(h-t)/h$. From (2.152), let $B(t) = r^2(h-t)^2\pi/h^2$ be the area of the circular region, e.g., πr_t^2. Thus, from (2.154),

$$V(S) = \int_0^h B(t)dt = \frac{r^2\pi}{h^2}\int_0^h \left(h^2 - 2ht + t^2\right)dt = \frac{r^2\pi}{h^2}\left[h^2t - ht^2 + \frac{t^3}{3}\right]_0^h = \frac{1}{3}hr^2\pi. \qquad \blacksquare$$

Example 2.30 *(Ghorpade and Limaye, 2018, Prop 8.5) The volume of a solid enclosed by an ellipsoid given by $(x^2/a^2) + (y^2/b^2) + (z^2/c^2) = 1$, where $a, b, c > 0$, is equal to $4\pi abc/3$. Letting $b = a$ and $c = a$, the volume of the spherical ball of radius a is equal to $4\pi a^3/3$.*

Proof: The given ellipsoid lies between the planes given by $x = -a$ and $x = a$. Also, for $s \in (-a, a)$, the area $A(s)$ of its slice

$$\left\{(s, y, z) \in \mathbb{R}^3 : \frac{y^2}{b^2} + \frac{z^2}{c^2} \leq 1 - \frac{s^2}{a^2}\right\}$$

by the plane given by $x = s$ is the area enclosed by the ellipse

$$\frac{y^2}{b^2\left(1 - (s^2/a^2)\right)} + \frac{z^2}{c^2\left(1 - (s^2/a^2)\right)} = 1,$$

and, hence, from (2.153),

$$A(s) = \pi\left(b\sqrt{1 - (s^2/a^2)}\right)\left(c\sqrt{1 - (s^2/a^2)}\right) = \pi bc\left(1 - \frac{s^2}{a^2}\right).$$

Thus the volume enclosed by the ellipsoid is equal to

$$\int_{-a}^a A(x)dx = \pi bc\int_{-a}^a \left(1 - \frac{x^2}{a^2}\right)dx = \pi bc\left(2a - \frac{2a^3}{3a^2}\right) = \frac{4}{3}\pi abc. \qquad \blacksquare$$

Example 2.31 *(Petrovic, Example 5.1.3) We wish to evaluate* $\int \frac{dx}{x-a}$. *If* $x > a$, *then* $\ln(x-a)$ *has derivative* $f(x) = 1/(x-a)$; *and from* (2.150), $F(x) = \int f(x)dx$ *is an indefinite integral of* f, *because* $F'(x) = f(x)$ *for all* $x > a$. *Thus,*

$$\int \frac{dx}{x-a} = \ln(x-a) + C, \text{ if } x > a. \tag{2.155}$$

On the other hand, if $x < a$, *then* $\ln(x-a)$ *is not defined. However,* $\ln(a-x)$ *is defined and differentiable, and its derivative is also* $1/(x-a)$, *so*

$$\int \frac{dx}{x-a} = \ln(a-x) + C, \text{ if } x < a. \tag{2.156}$$

These two formulae are usually combined to yield

$$\int \frac{dx}{x-a} = \ln|x-a| + C. \qquad \blacksquare$$

Augmenting the previous example a bit, consider the definite integral $I = \int_2^3 (x-1)^{-1} dx$, so $a = 1$, which, from (2.155) and the discussion just after (2.150), resolves to $\ln(x-1)|_2^3 = \ln 2$, recalling (2.86). Integral I can also be resolved with the substitution $u = x-1$, $du = dx$, so $I = \int_1^2 u^{-1} du = \ln 2$. For the case with $x < a$, let $a = 1$ and $I = \int_{-3}^{-2} (x-1)^{-1} dx = \ln(1-x)|_{-3}^{-2} = \ln 3 - \ln 4$, from (2.156). Alternatively, with $u = x-1$ and $du = dx$, $I = \int_{-4}^{-3} u^{-1} du$, which appears not to work, because we would get logs of negative numbers. But, from a plot of $1/u$, we know the area exists (and is negative). Let $v = -u$, $dv = -du$, so that $I = \int_4^3 (-v)^{-1}(-dv) = \int_4^3 v^{-1} dv = \ln 3 - \ln 4$.

Example 2.32 *To compute* $\int \frac{\sqrt{x}}{\sqrt{x}+1} dx$, *use the substitution* $\sqrt{x} = y$, *so* $dy/dx = \frac{1}{2\sqrt{x}} = \frac{1}{2y}$. *Note that*

$$y - 1 + \frac{1}{y+1} = \frac{(y-1)(y+1)}{(y+1)} + \frac{1}{y+1} = \frac{y^2 - 1 + 1}{(y+1)} = \frac{y^2}{y+1}.$$

Thus,

$$\begin{aligned} \int \frac{\sqrt{x}}{\sqrt{x}+1} dx &= \int \frac{y}{y+1} 2y dy = \int \frac{2y^2}{y+1} dy = \int \left[2(y-1) + \frac{2}{y+1} \right] dy \\ &= y^2 - 2y + 2\ln(y+1) = x - 2\sqrt{x} + 2\ln(\sqrt{x}+1) + C, \end{aligned}$$

having used the results in Example 2.31. $\blacksquare$

Example 2.33 *Observe that* $F(x) = e^{kx}/k + C$ *is a primitive of* $f(x) = e^{kx}$ *for* $k \in \mathbb{R} \setminus \{0\}$ *and any constant* $C \in \mathbb{R}$, *because, via the chain rule,* $dF(x)/dx = f(x)$. *Thus, from (2.143),*

$$\int_a^b f = F(b) - F(a) = k^{-1}(e^{kb} - e^{ka}). \tag{2.157}$$

See (2.170) *for an example that is associated with the exponential distribution.* $\blacksquare$

Example 2.34 *Let* $I(x) = \int_0^{x^2} e^{-t} dt = 1 - e^{-x^2}$, *so that* $I'(x) = \frac{d}{dx}\left(1 - e^{-x^2}\right) = 2xe^{-x^2}$. *Alternatively, let* $G(y) = \int_0^y e^{-t} dt$, *so that* $I(x) = G(x^2) = G(f(x))$, *where* $f(x) = x^2$. *From* (2.144), $G'(y) = e^{-y}$, *and from the chain rule,*

$$I'(x) = G'(f(x)) f'(x) = e^{-x^2} \cdot 2x,$$

as before, but without having to actually evaluate $I(x)$. $\blacksquare$

Example 2.35 *Recall (2.140), the Mean Value Theorem for Integrals: For $f, p \in \mathcal{C}^0(I)$, $I = [a,b]$, with p nonnegative, $\exists c \in I$ such that $\int_a^b f(x)p(x)\,dx = f(c)\int_a^b p(x)\,dx$. The FTC allows an easy proof of this. As f is integrable from (2.130), let $F(x) = \int_a^x f$. From (2.145), F is continuous; and from (2.146), $F'(x) = f(x)$ for all $x \in I$, i.e., F is differentiable on I. The Mean Value Theorem (2.63) thus implies $\exists c \in I$ such that*

$$\frac{F(b) - F(a)}{b-a} = F'(c),$$

i.e., $\int_a^b f(x)dx = F(b) - F(a) = F'(c)(b-a) = f(c)(b-a)$. ■

The simple technique of *integration by parts* can be invaluable in many situations.

Theorem: Let $f, g \in \mathcal{R}[a,b]$ with primitives F and G, respectively. Then, in the definite integral case,

$$\begin{aligned}\int_a^b F(t)\,g(t)\,dt &= F(b)\,G(b) - F(a)\,G(a) - \int_a^b f(t)\,G(t)\,dt \\ &= FG\Big|_a^b - \int_a^b f(t)\,G(t)\,dt,\end{aligned} \tag{2.158}$$

while for the indefinite integral,

$$\int F(t)\,g(t)\,dt = FG - \int f(t)\,G(t)\,dt. \tag{2.159}$$

Proof: Use the product rule to get $(FG)' = F'G + FG' = fG + Fg$. Integrating both sides of this and using FTC (2.143) for the lhs gives, from the linearity property of the Riemann integral (2.134),

$$F(b)\,G(b) - F(a)\,G(a) = \int_a^b [f(t)\,G(t) + F(t)\,g(t)]\,dt,$$

and

$$FG = \int [f(t)\,G(t) + F(t)\,g(t)]\,dt.$$

These are equivalent to (2.158) and (2.159), respectively.

Throughout, we will use the popular notation

$$\int_a^b u\,dv = uv\Big|_a^b - \int_a^b v\,du \qquad \text{or} \qquad \int u\,dv = uv - \int v\,du, \tag{2.160}$$

where $uv|_a^b := u(b)\,v(b) - u(a)\,v(a)$.

Example 2.36 *Let $f(x) = x\exp(-x)$ and $I = \int f$. Using the latter equation in (2.160) with $u = x$ and $dv = \exp(-x)dx$, we obtain*

$$I = uv - \int v du = -x\exp(-x) + \int \exp(-x)dx = -x\exp(-x) - \exp(-x).$$

Differentiating the rhs gives $(-x)(-\exp(-x)) + (-1)\exp(-x) + \exp(-x) = f(x)$. ■

Example 2.37 *To compute* $\int \ln x\,dx$, *let* $u = \ln x$, $du = dx/x$, $dv = dx$, $v = x$, *so that*

$$\int \ln x\,dx = x\ln x - \int x(1/x)dx = x\ln x - x.$$

In the definite integral case, let $f(t) = \ln t$ *for* $t \in (0,1]$. *For* $x \in (0,1]$,

$$\int_x^1 f(t)dt = (t\ln t - t)\big|_x^1 = x - 1 - x\ln x.$$

We need to appeal to (2.168) *for improper integrals. Since* $x\ln x \to 0$ *as* $x \to 0^+$, *we see that* $\int_{0<t\leq 1} \ln t dt$ *is convergent and its value is* -1. ■

Example 2.38 *Applying integration by parts to* $\int_0^1 x^r (\ln x)^r\,dx$ *for* $r \in \mathbb{N}$ *with* $u = (\ln x)^r$ *and* $dv = x^r dx$ *(so that* $v = x^{r+1}/(r+1)$ *and* $du = r(\ln x)^{r-1}x^{-1}dx$*) gives*

$$\int_0^1 x^r (\ln x)^r\,dx = (\ln x)^r \frac{x^{r+1}}{r+1}\bigg|_0^1 - \int_0^1 \frac{x^{r+1}}{r+1}\frac{r}{x}(\ln x)^{r-1}\,dx = -\frac{r}{r+1}\int_0^1 x^r(\ln x)^{r-1}\,dx.$$

Repeating this "in a feast of integration by parts" (Havil, 2003, p. 44) leads to

$$\int_0^1 x^r (\ln x)^r\,dx = (-1)^r \frac{r!}{(r+1)^{r+1}}, \quad r \in \mathbb{N}, \tag{2.161}$$

which is used in Example 2.83 below. ■

Example 2.39 *(Bóna and Shabanov, p. 30) To compute* $\int e^x \cos x\,dx$, *set* $u = \cos x$ *and* $dv = e^x\,dx$. *Then* $du = -\sin x\,dx$, $v = e^x$, *and*

$$\int e^x \cos x\,dx = e^x\cos x + \int e^x \sin x\,dx. \tag{2.162}$$

So we could solve our problem if we could compute the integral $\int e^x \sin x\,dx$. *We can do that by applying the technique of integration by parts again, obtaining*

$$\int e^x \sin x\,dx = e^x\sin x - \int e^x\cos x\,dx. \tag{2.163}$$

Finally, note that (2.162) *and* (2.163) *is a system of equations with unknowns* $\int e^x\cos x\,dx$ *and* $\int e^x \sin x\,dx$. *By adding these two equations, we get*

$$\int e^x\cos x\,dx = e^x(\cos x + \sin x) - \int e^x\cos x\,dx$$

or

$$\int e^x\cos x\,dx = \frac{e^x}{2}(\cos x + \sin x).$$

Note that substituting the obtained expression for $\int e^x\cos x\,dx$ *into* (2.163), *we get a formula for* $\int e^x\sin x\,dx$, *namely,*

$$\int e^x \sin x\,dx = \frac{e^x}{2}(\sin x - \cos x). \qquad ■$$

Example 2.40 *(Petrovic, Example 5.1.6) To compute* $\int \frac{dx}{x \ln x}$, *use* $u = \ln x$, *so* $du = dx/x$, *and, recalling Example 2.31,*

$$\int \frac{dx}{x \ln x} = \int \frac{du}{u} = \ln |u| + C = \ln |\ln x| + C. \qquad \blacksquare$$

Example 2.41 *Recall (2.87) and (2.89). The natural logarithm is often defined as* $\ln x = \int_1^x t^{-1} dt$, *for* $x > 0$, *from which its most important properties follow, such as* $\ln 1 = 0$ *and* $\ln(xy) = \ln x + \ln y$. *For the latter, let* $u = t/x$, *so that* $t = xu$ *and* $dt = xdu$. *Then*

$$\int_x^{xy} t^{-1} dt = \int_1^y \frac{1}{u} du.$$

Thus, using domain additivity (2.138),

$$\ln(xy) = \int_1^{xy} t^{-1}\, dt = \int_1^x t^{-1}\, dt + \int_x^{xy} t^{-1}\, dt = \int_1^x t^{-1}\, dt + \int_1^y t^{-1}\, dt = \ln x + \ln y.$$

Similarly, $\ln(x/y) = \ln x - \ln y$.

Now let $y \in \mathbb{Z}$. *If we use the definition of the log from §2.2.4 and its various properties there, we obtain*

$$\frac{d}{dx} \ln x^y = \frac{1}{x^y} \frac{dx^y}{dx} = \frac{1}{x^y} y x^{y-1} = \frac{y}{x}$$

and

$$\frac{d}{dx} y \ln x = \frac{y}{x},$$

so that $\ln x^y$ *and* $y \ln x$ *have the same first derivative, and thus differ by a constant,* $C = \ln x^y - y \ln x$. *With* $y = 0$, $C = \ln 1 = 0$ *from (2.86), arriving at what we already know from (2.87) and (2.89), namely* $\ln x^y = y \ln x$.

What we wish to do is repeat this exercise, but using the integral definition, $\ln x = \int_1^x t^{-1} dt$. *From the integral representation,* $\ln x^y = \int_1^{x^y} t^{-1} dt$. *Thus, the FTC and the chain rule imply that*

$$\frac{d}{dx} \ln x^y = \frac{1}{x^y} \frac{dx^y}{dx} = \frac{1}{x^y} y x^{y-1} = \frac{y}{x}.$$

Likewise,

$$\frac{d}{dx} y \ln x = y \frac{d}{dx} \int_1^x t^{-1} dt = \frac{y}{x},$$

so that, as before, $\ln x^y$ *and* $y \ln x$ *have the same first derivative, and thus differ by a constant,* $C = \ln x^y - y \ln x$. *With* $y = 0$, *and using the first definition in (2.139),* $C = \ln 1 - 0 = 0$, *so that* $\ln x^y = y \ln x$. *This is easily extended to* $y \in \mathbb{Q}$ *(see, e.g., Protter and Morrey, 1991, p. 118). The extension to* $y \in \mathbb{R}$ *follows by* defining $x^r = \exp(r \ln x)$, *as was given in §2.2.4.* $\blacksquare$

Example 2.42 *Let* $b, c \in \mathbb{R}$ *and* $T \in \mathbb{R}_{>0}$. *Recall (2.51), i.e.,* $\cos(-x) = \cos(x)$ *and* $\sin(-x) = -\sin(x)$. *Letting* $u = -t$,

$$\begin{aligned} A &= \int_{-T}^{T} t^{-1} \sin(bt) \sin(ct) dt \\ &= \int_{T}^{-T} (-u^{-1}) \sin(-bu) \sin(-cu)(-1) du = -\int_{-T}^{T} u^{-1} \sin(bu) \sin(cu) du = -A, \end{aligned}$$

so that $A = 0$. *Similarly,*

$$\begin{aligned} B &= \int_0^T t^{-1}\sin(bt)\cos(ct)dt \\ &= \int_0^{-T}(-u^{-1})\sin(-bu)\cos(-cu)(-1)du = \int_{-T}^0 u^{-1}\sin(bu)\cos(cu)du, \end{aligned}$$

so that

$$\int_{-T}^T t^{-1}\sin(bt)\cos(ct)dt = 2\int_0^T t^{-1}\sin(bt)\cos(ct)dt.$$

These results are useful when working with characteristic functions of random variables. ■

Another useful integration technique when working with ratios of polynomials is *partial fraction decomposition.* Nice presentations can be found in, e.g., Petrovic, Advanced Calculus: Theory and Practice, 2nd ed., 2020; and Bóna and Shabanov, Concepts in Calculus II, 2012. We give one such example, from the latter book, to illustrate the idea; and the reader is encouraged to inspect the two aforementioned books, and others, for more detail and further examples.

Example 2.43 *(Bóna and Shabanov, p. 41) To compute* $\int \frac{1}{x^2+3x+2}\,dx$, *note that* $x^2+3x+2 = (x+1)(x+2)$. *Using that observation, we are looking for real numbers* A *and* B *such that*

$$\frac{1}{x^2+3x+2} = \frac{A}{x+1} + \frac{B}{x+2} \tag{2.164}$$

as functions, that is, such that (2.164) *holds for all real numbers* x. *Multiplying both sides by* x^2+3x+2, *we get*

$$1 = A(x+2) + B(x+1). \tag{2.165}$$

If (2.165) *holds for all real numbers* x, *it must hold for* $x = -1$ *and* $x = -2$ *as well. However, if* $x = -1$, *then* (2.165) *reduces to* $1 = A$, *and if* $x = -2$, *then* (2.165) *reduces to* $1 = -B$. *So we conclude that* $A = 1$ *and* $B = -1$ *are the numbers we wanted to find. It is now easy to compute the requested integral as follows:*

$$\int \frac{dx}{x^2+3x+2} = \int \frac{dx}{x+1} - \int \frac{dx}{x+2} = \ln(x+1) - \ln(x+2) + C,$$

having used the results in Example 2.31. ■

Example 2.44 *We will require the trigonometric secant function, given by* $\sec\theta = 1/\cos\theta$. *We wish to show the related integrals*

$$\int \sec y\; dy = \ln|\sec y + \tan y| \quad \textit{and} \quad \int \left(x^2-1\right)^{-1/2} dx = \ln\left|x + \sqrt{x^2-1}\right|.$$

Before proceeding, we use symbolic computing (Maple in this case) to check (apparent) equivalence of power series expansions of the first integral. Indeed, the two expansions appear equal, namely, and having used termwise integration,

$$\begin{aligned} \sec y &= 1 + \frac{1}{2}y^2 + \frac{5}{24}y^4 + \frac{61}{720}y^6 + \frac{277}{8064}y^8 + \cdots \\ \int \sec y dy &= y + \frac{1}{6}y^3 + \frac{1}{24}y^5 + \frac{61}{5040}y^7 + \frac{277}{72\,576}y^9 + \cdots = \ln\left(\sec y + \tan y\right). \end{aligned}$$

Now note the simple identity

$$\tan^2\theta = \frac{\sin^2\theta}{\cos^2\theta} = \frac{1-\cos^2\theta}{\cos^2\theta} = \sec^2\theta - 1, \tag{2.166}$$

which implies, for $y \in [0, \pi/2)$*, that* $\tan\theta = \sqrt{\sec^2\theta - 1}$*, so that the substitution* $x = \sec y$ *could be useful. Note from the quotient rule for derivatives that*

$$\frac{dx}{dy}\sec y = \frac{\sin y}{\cos^2 y} = \tan y \cdot \sec y, \quad \textit{i.e.,} \quad dx = \tan y \cdot \sec y \; dy.$$

Using the (for now, and subsequently proven) result for the first integral above; the suggested substitution $x = \sec y$*; and that, from (2.166),* $\sqrt{x^2-1} = \sqrt{\sec^2 y - 1} = \tan y$*, we get*

$$\int \frac{dx}{\sqrt{x^2-1}} = \int \frac{\tan y \cdot \sec y}{\tan y} dy = \int \sec y dy = \ln|\sec y + \tan y| \; = \ln\left|x + \sqrt{x^2-1}\right|.$$

As a "numerical confirmation", indeed,

$$\int_1^3 \left(x^2-1\right)^{-1/2} dx \approx 1.7627 \approx \ln\left|x + \sqrt{x^2-1}\right|\Big|_1^3.$$

What remains is to prove that $\int \sec y \; dy = \ln|\sec y + \tan y|$*. It turns out that this integral has quite some history, and for which there is a highly informative and detailed Wikipedia web page, Integral_of_the_secant_function, from which we obtain the following derivation.* ■

Example 2.45 *Here we resolve the interesting integral in Example 2.44, using, as said, the Wikipedia entry. It turns out that there are several equivalent expressions for the integral, the common first three of which are (and leaving off the "constant of integration" C)*

$$\int \sec\theta d\theta = \frac{1}{2}\ln\frac{1+\sin\theta}{1-\sin\theta} = \ln|\sec\theta + \tan\theta| = \ln\left|\tan\left(\frac{\theta}{2} + \frac{\pi}{4}\right)\right|.$$

We first prove that these are equivalent, because

$$\sqrt{\frac{1+\sin\theta}{1-\sin\theta}} = |\sec\theta + \tan\theta| = \left|\tan\left(\frac{\theta}{2} + \frac{\pi}{4}\right)\right|. \tag{2.167}$$

Proof of (2.167): We can separately apply the so-called tangent half-angle substitution $t = \tan\frac{1}{2}\theta$ *to each of the three forms, and show them equivalent to the same expression in terms of* t*. Under this substitution,* $\cos\theta = (1-t^2)/(1+t^2)$ *and* $\sin\theta = 2t/(1+t^2)$*. First,*

$$\sqrt{\frac{1+\sin\theta}{1-\sin\theta}} = \sqrt{\frac{1+\frac{2t}{1+t^2}}{1-\frac{2t}{1+t^2}}} = \sqrt{\frac{1+t^2+2t}{1+t^2-2t}} = \sqrt{\frac{(1+t)^2}{(1-t)^2}} = \left|\frac{1+t}{1-t}\right|.$$

Second,

$$|\sec\theta + \tan\theta| = \left|\frac{1}{\cos\theta} + \frac{\sin\theta}{\cos\theta}\right| = \left|\frac{1+t^2}{1-t^2} + \frac{2t}{1-t^2}\right| = \left|\frac{(1+t)^2}{(1+t)(1-t)}\right| = \left|\frac{1+t}{1-t}\right|.$$

Third, using the tangent addition identity $\tan(\phi+\psi) = (\tan\phi + \tan\psi)/(1 - \tan\phi\tan\psi)$*,*

$$\left|\tan\left(\frac{\theta}{2} + \frac{\pi}{4}\right)\right| = \left|\frac{\tan\frac{1}{2}\theta + \tan\frac{1}{4}\pi}{1 - \tan\frac{1}{2}\theta\tan\frac{1}{4}\pi}\right| = \left|\frac{t+1}{1-t\cdot 1}\right| = \left|\frac{1+t}{1-t}\right|.$$

Thus, all three expressions describe the same quantity.

There are also several approaches to resolving the integral, all shown in detail in the Wikipedia page, and of which we show one, the so-called Barrow's approach from the year 1670, using a partial fraction decomposition and the results we have in Example 2.31.

Proof (Barrow): Write

$$\int \sec\theta d\theta = \int \frac{1}{\cos\theta} d\theta = \int \frac{\cos\theta}{\cos^2\theta} d\theta = \int \frac{\cos\theta}{1-\sin^2\theta} d\theta.$$

Substituting $u = \sin\theta, du = \cos\theta d\theta$, *reduces the integral to*

$$\begin{aligned}\int \frac{1}{1-u^2} du &= \int \frac{1}{(1+u)(1-u)} du = \int \frac{1}{2}\left(\frac{1}{1+u} + \frac{1}{1-u}\right) du \\ &= \frac{1}{2}(\ln|1+u| - \ln|1-u|) + C = \frac{1}{2}\ln\left|\frac{1+u}{1-u}\right| + C.\end{aligned}$$

Therefore,

$$\int \sec\theta d\theta = \frac{1}{2}\ln\frac{1+\sin\theta}{1-\sin\theta} + C.$$

Taking the absolute value is not necessary because $1+\sin\theta$ *and* $1-\sin\theta$ *are always non-negative for real values of* θ. ■

2.4.3 Improper Integrals

Recall that the Riemann integral is designed for bounded functions on a closed, bounded interval domain. An extension, credited to Cauchy, is to let $f : (a, b] \to \mathbb{R}$ such that, $\forall c \in (a, b)$, f is integrable on $[c, b]$, and define

$$\int_a^b f = \lim_{c\to a^+} \int_c^b f. \tag{2.168}$$

A similar definition holds when the limit is taken at the upper boundary. These are termed *improper integrals (of the second kind)*. If the limit in (2.168) exists, then $\int_a^b f$ is *convergent*, otherwise *divergent*.

Example 2.46 *(Stoll, 2001, p. 241) Let* $f(x) = x^{-1/2}$, *for* $x \in (0, 1]$. *Then*

$$\int_0^1 f = \lim_{c\to 0^+} \int_c^1 x^{-1/2} dx = \lim_{c\to 0^+} \left(2 - 2\sqrt{c}\right) = 2,$$

and $\int_0^1 f$ *is convergent. As* $f \in \mathcal{R}[c, 1]$ *for* $c \in (0, 1)$, *(2.132) implies that* $f^2 \in \mathcal{R}[c, 1]$. *But that does* not *imply that the improper integral* $\int_0^1 f^2$ *is convergent. Indeed, recalling the integral definition of log from Example 2.41,*

$$\lim_{c\to 0^+} \int_c^1 x^{-1} dx = -\lim_{c\to 0^+} \ln c = \infty,$$

i.e., $\int_0^1 f^2$ *is divergent.* ■

Integrals can also be taken over infinite intervals, i.e., $\int_a^\infty f(x)\,dx$, $\int_{-\infty}^b f(x)\,dx$ and $\int_{-\infty}^\infty f(x)\,dx$; these are (also) referred to as *improper integrals (of the first kind)*. If function f is defined on $(-\infty,\infty)$, then $\int_{-\infty}^\infty f(x)\,dx$ is defined by

$$\int_{-\infty}^{\infty} f(x)\,dx = \lim_{a\to-\infty}\int_a^t f(x)\,dx + \lim_{b\to\infty}\int_t^b f(x)\,dx, \tag{2.169}$$

for any point $t\in\mathbb{R}$, when both limits on the rhs exist. An example that we will use often is

$$\int_0^\infty e^{-u}du = \lim_{b\to\infty}\int_0^b e^{-u}du = \lim_{b\to\infty}\left(1-e^{-b}\right) = 1, \tag{2.170}$$

having used (2.157) with $k=-1$.

We consider a few examples first, and then return to some basic theory.

Example 2.47 *The unpleasant looking integral*

$$\int_0^1 \frac{e^{-((1/v)-1)}}{v^2}dv$$

is easily handled by using the substitution $u=(1/v)-1$, *so that* $v=1/(1+u)$ *and* $dv=-(1+u)^{-2}\,du$. *Thus,*

$$\int_0^1 \frac{e^{-((1/v)-1)}}{v^2}dv = -\int_\infty^0 \frac{e^{-u}}{(1+u)^{-2}}(1+u)^{-2}\,du = \int_0^\infty e^{-u}du = 1,$$

from (2.170). ■

Example 2.48 *Applying integration by parts to* $\int e^{at}\cos(bt)\,dt$ *with* $u=e^{at}$ *and* $dv=\cos(bt)\,dt$ *(so that* $du=ae^{at}dt$ *and* $v=(\sin bt)/b$*) gives*

$$\int e^{at}\cos(bt)\,dt = e^{at}\frac{\sin bt}{b} - \frac{a}{b}\int e^{at}\sin(bt)\,dt.$$

Similarly,

$$\int e^{at}\sin(bt)\,dt = -e^{at}\frac{\cos bt}{b} + \frac{a}{b}\int e^{at}(\cos bt)\,dt,$$

so that

$$\int e^{at}\cos(bt)\,dt = e^{at}\frac{\sin bt}{b} + e^{at}\frac{a\cos bt}{b^2} - \frac{a^2}{b^2}\int e^{at}(\cos bt)\,dt,$$

or

$$\int e^{at}\cos(bt)\,dt = \frac{e^{at}}{a^2+b^2}(a\cos bt + b\sin bt). \tag{2.171}$$

This is given, for example, in Abramowitz and Stegun (1972, eq. 4.3.137). In the definite integral case over the positive real line, for $a=-1$ *and* $b=s$, *it is easy to confirm that (2.171) reduces to*

$$\int_0^\infty e^{-t}\cos(st)\,dt = \frac{1}{1+s^2}. \tag{2.172}$$

A similar derivation confirms that

$$\int e^{at}\sin(bt)\,dt = \frac{e^{at}}{a^2+b^2}(a\sin bt - b\cos bt),$$

with special case

$$\int_0^\infty e^{-t}\sin(st)\,dt = \frac{s}{1+s^2}, \tag{2.173}$$

for $a=-1$ and $b=s$. We will require (2.172) and (2.173) below.

Taking $a=-s$ and $b=1$ yield

$$\int_0^\infty e^{-st}\cos(t)dt = \frac{s}{s^2+1}, \quad \int_0^\infty e^{-st}\sin(t)dt = \frac{1}{s^2+1}, \tag{2.174}$$

which also can be derived using Laplace transforms and basic complex analysis; see, e.g., Paolella, Intermediate Probability, Example 1.17. ■

Example 2.49 *The integral $\int_{-\infty}^\infty \exp(-x^2)\,dx$ is important, arising in conjunction with the Gaussian distribution (they are related by substituting $y=x/\sqrt{2}$). In this example, we only wish to verify that it is convergent. Its value will be computed in Examples 2.55 and 5.20 below.*

Via symmetry, it suffices to study $\int_0^\infty \exp(-x^2)\,dx$. Let $f(x)=e^{-x^2}$, $x\in\mathbb{R}_{\geq 0}$. As f is bounded and monotone on $[0,k]$ for all $k\in\mathbb{R}_{>0}$, it follows from (2.129) that $\int_0^k f$ exists. Alternatively, continuity of composite functions (2.25), and (2.130), also imply its existence. Thus, for examining the limit as $k\to\infty$, it suffices to consider $\int_1^k f$.

To proceed, we require the Taylor series expansion of $\exp(x)$. The general Taylor expansion is given in (2.282), and that for the exponential function is given in (2.241). From the latter, $e^{x^2}=1+x^2+\frac{1}{2}(x^2)^2+\cdots > x^2$ (for $x\in\mathbb{R}$), and it follows that, for $x>0$,

$$e^{x^2} > x^2 \Rightarrow x^2 > 2\ln x \Rightarrow -x^2 < -2\ln x \Rightarrow e^{-x^2} < x^{-2},$$

recalling that $-2\ln x = \ln(x^{-2})$ from (2.89). Thus, from integral monotonicity (2.133),

$$\int_1^k e^{-x^2}dx < \int_1^k x^{-2}dx = \frac{k-1}{k} < 1, \quad \forall k>1,$$

and $\int_1^\infty e^{-x^2}dx$ is convergent. Alternatively, for $x>1$, and that exp is strictly increasing,

$$x^2 > x \Rightarrow -x^2 < -x \Rightarrow e^{-x^2} < e^{-x},$$

so that, from integral monotonicity (2.133), result (2.170), and that $\forall x\in\mathbb{R}$, $\exp(x)>0$,

$$\int_1^k e^{-x^2}\,dx < \int_1^k e^{-x}\,dx = e^{-1}-e^{-k} < e^{-1} \approx 0.367879, \quad \forall k>1. \tag{2.175}$$

Thus, $\int_1^\infty e^{-x^2}dx$ and, hence, $\int_0^\infty e^{-x^2}dx$ are convergent.

To see how close (2.175) is to the true value, use of (2.138), the result in Example 2.55, and numeric integration gives

$$\int_1^\infty e^{-x^2}dx = \int_0^\infty e^{-x^2}dx - \int_0^1 e^{-x^2}dx \approx \frac{\sqrt{\pi}}{2} - 0.746824 \approx 0.1394.$$ ■

Example 2.50 *To show that*

$$I := \int_0^\infty (1+t)\,e^{-t}\cos(st)\,dt = \frac{2}{(1+s^2)^2}, \tag{2.176}$$

let $C = \int_0^\infty te^{-t}\cos(st)\,dt$ *and* $S = \int_0^\infty te^{-t}\sin(st)\,dt$. *Set* $u = te^{-t}$ *and* $dv = \cos(st)\,dt$ *so that* $du = (1-t)\,e^{-t}dt$ *and* $v = (\sin st)/s$. *Then, from (2.173),*

$$\begin{aligned} C &= te^{-t}\frac{\sin st}{s}\bigg|_{t=0}^{\infty} - \int_0^\infty \frac{\sin st}{s}(1-t)\,e^{-t}dt \\ &= 0 \quad -\frac{1}{s}\int_0^\infty e^{-t}\sin(st)\,dt + \frac{1}{s}\int_0^\infty te^{-t}\sin(st)\,dt = -\frac{1}{1+s^2} + \frac{S}{s}. \end{aligned}$$

Similarly, with $dv = \sin(st)\,dt$, $v = -(\cos st)/s$ *and using (2.172),*

$$\begin{aligned} S &= -te^{-t}\frac{\cos st}{s}\bigg|_{t=0}^{\infty} + \int_0^\infty \frac{\cos st}{s}(1-t)\,e^{-t}dt \\ &= 0 \quad +\frac{1}{s}\int_0^\infty e^{-t}\cos(st)\,dt - \frac{1}{s}\int_0^\infty te^{-t}\cos(st)\,dt = \frac{1}{s(1+s^2)} - \frac{C}{s}. \end{aligned}$$

Combining these yields

$$C = \int_0^\infty te^{-t}\cos(st)\,dt = \frac{1-s^2}{(1+s^2)^2}, \tag{2.177}$$

so that, from (2.172) and (2.177),

$$I = \frac{1}{1+s^2} + \frac{1-s^2}{(1+s^2)^2} = \frac{2}{(1+s^2)^2},$$

which is (2.176). ■

Example 2.51 *Let* $f(x) = x/(1+x^2)$. *Then, using the substitution* $u = 1+x^2$, *a straightforward calculation yields*

$$\int_0^c \frac{x}{1+x^2}\,dx = \frac{1}{2}\ln\left(1+c^2\right) \quad \textit{and} \quad \int_{-c}^0 \frac{x}{1+x^2}\,dx = -\frac{1}{2}\ln\left(1+c^2\right),$$

which implies that

$$\lim_{c\to\infty}\int_{-c}^{c}\frac{x}{1+x^2}\,dx = \lim_{c\to\infty} 0 = 0. \tag{2.178}$$

This would seem to imply that

$$0 = \lim_{c\to\infty}\int_{-c}^{c}\frac{x}{1+x^2}\,dx \stackrel{?}{=} \int_{-\infty}^{\infty}\frac{x}{1+x^2}\,dx,$$

but the second equality is not true, because the limits in (2.169) do not exist. In (2.178), the order is conveniently chosen so that positive and negative terms precisely cancel, resulting in zero. An application of this is that the expectation of a Cauchy random variable does not exist. ■

Remarks:
(a) A similar calculation shows that, for $c > 0$ and $k > 0$,

$$\lim_{c\to\infty}\int_{-c}^{kc}\frac{x}{1+x^2}\,dx = \ln k.$$

This expression could also be used for evaluating $\int_{-\infty}^{\infty} f(x)\,dx$, but results in a different value for each k. Thus, it also shows that $\int_{-\infty}^{\infty} f(x)\,dx$ does not exist.

(b) Notice that $f(x) = (1+x^2)^{-1}$ is an *even function*, i.e., it satisfies $f(-x) = f(x)$ for all x (or is symmetric about zero). In this case, f is continuous for all x, so that, for any finite $c > 0$, $\int_0^c f(x)dx = \int_{-c}^0 f(x)dx$. On the other hand, $g(x) = x$ is an *odd function*, i.e., satisfies $g(-x) = -g(x)$, and, as g is continuous, for any finite $c > 0$, $\int_0^c g(x)dx = -\int_{-c}^0 g(x)dx$. Finally, as $h(x) = f(x)g(x)$ is also odd, $\int_{-c}^c h(x)dx = 0$. Thus, the result in (2.178) could have been immediately determined.

(c) The integral $\int_a^{\infty} \cos x\,dx$ also does not exist, because $\sin x$ does not have a limit as $x \to \infty$. Notice, however, that, for any value $t > 0$, the integral $\int_a^t \cos x\,dx$ is bounded. This shows that, if $\int_a^{\infty} f(x)dx$ does not exist, then it is not necessarily true that $\int_a^t f(x)dx$ increases as $t \to \infty$. ■

Example 2.52 *(Example 2.51 cont.) We have seen that* $\int_{-\infty}^{\infty} x\,(1+x^2)^{-1}\,dx$ *does not exist, but, for any* $z \in \mathbb{R}$,

$$I = \int_{-\infty}^{\infty} f(x)\,dx := \int_{-\infty}^{\infty} \left(\frac{x}{1+x^2} - \frac{x}{1+(z-x)^2} \right) dx = -\pi z \tag{2.179}$$

exists. We require (2.179) *below in* (2.182).

Recall (2.62), *i.e.,* $(d/dx)\arctan(z-x) = -\left[1+(z-x)^2\right]^{-1}$. *Let*

$$F(x) = \frac{1}{2}\ln\left(\frac{1+x^2}{1+(z-x)^2}\right) + z\arctan(z-x),$$

so that (having used Maple to verify the second to last equality),

$$\begin{aligned} F'(x) &= \frac{1}{2}\frac{1+(z-x)^2}{1+x^2}\frac{\left(1+(z-x)^2\right)2x + \left(1+x^2\right)2(z-x)}{\left(1+(z-x)^2\right)^2} - \frac{z}{1+(z-x)^2} \\ &= \frac{x}{1+x^2} - \frac{x}{1+(z-x)^2} = f(x), \end{aligned}$$

i.e., from (2.150), F *is an indefinite integral for* f. *We still need to address the bounds on the integral, with* (2.179) *being an improper integral. For fixed* z, *and using* (2.25),

$$\lim_{x\to\pm\infty} \ln\left(\frac{1+x^2}{1+(z-x)^2}\right) = \ln\left(\lim_{x\to\pm\infty} \frac{1+x^2}{1+(z-x)^2}\right) = \ln 1 = 0.$$

From, e.g., a graph of $\tan(x)$ *for* $-\pi/2 < x < \pi/2$, *and that its inverse,* arctan *is strictly increasing (see, e.g., the discussion following* (2.185)), *we see that, for fixed* z,

$$\lim_{x\to\infty} \arctan(z-x) = -\frac{\pi}{2} \qquad \textit{and} \qquad \lim_{x\to-\infty} \arctan(z-x) = \frac{\pi}{2}, \tag{2.180}$$

from which the value of $-\pi z$ *for* (2.179) *follows.* ■

Example 2.53 *(Example 2.52 cont.) Consider the integral*

$$I(s) = \int_{-\infty}^{\infty} \frac{1}{1+x^2}\,\frac{1}{1+(s-x)^2}\,dx.$$

To resolve this, the first step is to use a partial fraction decomposition *for the integrand,*

$$\frac{1}{1+x^2}\,\frac{1}{1+(s-x)^2} = \frac{Ax}{1+x^2} + \frac{B}{1+x^2} - \frac{Cx}{1+(s-x)^2} + \frac{D}{1+(s-x)^2},$$

where

$$A = \frac{2}{sR}, \quad B = \frac{1}{R}, \quad C = A, \quad D = \frac{3}{R},$$

and $R = s^2 + 4$*, which can be easily verified with symbolic math software. From linearity* (2.134)*, we can integrate each of the four above pieces. With substitution* $u = s - x$*,* $du = -dx$*,* $x \to \infty \Rightarrow u \to -\infty$*, and* $x \to -\infty \Rightarrow u \to \infty$*,*

$$\int_{-\infty}^{\infty} \frac{dx}{1+(s-x)^2} = \int_{-\infty}^{\infty} \frac{du}{1+u^2}.$$

From (2.61)*, namely, for* $g(y) = \arctan(y)$*,* $g'(y) = 1/(1+y^2)$*, we see that* $g(y)$ *is an indefinite integral of* $1/(1+y^2)$*. Thus, from* (2.150)*,* $\int du/(1+u^2) = \arctan(u)$*, and, from* (2.180) *with* $z = 0$*,*

$$\int_{-\infty}^{\infty} \frac{dx}{1+(s-x)^2} = \int_{-\infty}^{\infty} \frac{du}{1+u^2} = \lim_{u\to\infty} \arctan u - \lim_{u\to-\infty} \arctan u = \pi.$$

Thus, integration of the B *and* D *terms gives*

$$\int_{-\infty}^{\infty} \left(\frac{B}{1+x^2} + \frac{D}{1+(s-x)^2} \right) dx = \pi(B+D) = \frac{4\pi}{R} = \frac{4\pi}{s^2+4}. \tag{2.181}$$

For the remaining two terms, use of (2.179) *leads to*

$$\begin{aligned} \int_{-\infty}^{\infty} \left(\frac{Ax}{1+x^2} - \frac{Cx}{1+(s-x)^2} \right) dx &= A \int_{-\infty}^{\infty} \left(\frac{x}{1+x^2} - \frac{x}{1+(s-x)^2} \right) dx \\ &= -A\pi s = \frac{2}{sR}\pi s = -\frac{2\pi}{s^2+4}, \end{aligned} \tag{2.182}$$

so that, adding (2.181) *and* (2.182)*,*

$$I(s) = \frac{2\pi}{s^2+4} = \frac{\pi}{2}\frac{1}{1+(s/2)^2}.$$

This result is required for determining the density of the sum of two independent standard Cauchy random variables via the convolution formula. ■

The indefinite integral $\int r^{-1}\sin r\,dr$ is termed the Si function. In preparation for the next example, we wish to consider $F(x) := \int_0^x r^{-1}\sin r\;dr$ for $x > 0$, and show that $\mathrm{Si}(0) = \lim_{x\to 0^+} F(x) = 0$. From (2.56), $\lim_{r\to 0} \sin(r)/r = 1$, so, if we define the integrand as $r^{-1}\sin r$, for $r > 0$, and 0, for $r = 0$, then the integrand is defined, bounded, and continuous on any closed interval $[0, x]$, $x > 0$. Thus, from the lhs of (2.139), $F(0) = 0$. We play around with some other ways to determine this.

For $r > 0$, $r^{-1}\sin r \le |r^{-1}\sin r| \le |r^{-1}| = r^{-1}$, so monotonicity of the integral implies, for $x > 0$,

$$F(x) = \int_0^x r^{-1}\sin r\;dr \le \int_0^x r^{-1} dr.$$

For $0 < a, b < 1$, and recalling the integral representation of log in Example 2.41,

$$\int_a^b \frac{dx}{x} = \int_a^1 \frac{dx}{x} - \int_b^1 \frac{dx}{x} = \int_1^b \frac{dx}{x} - \int_1^a \frac{dx}{x} = \ln b - \ln a,$$

and, for $b = ka$ for $k > 1$,

$$\lim_{k\to 1^+} \lim_{a\to 0^+} (\ln b - \ln a) = \lim_{k\to 1^+} \lim_{a\to 0^+} (\ln ka - \ln a) = \lim_{k\to 1^+} \lim_{a\to 0^+} (\ln k + \ln a - \ln a) = \lim_{k\to 1^+} \ln k = 0.$$

Clearly, for $0 \le x \le \pi/2$, $\int_0^x r^{-1} \sin r \; dr \ge 0$, so the Squeeze Theorem implies that $\lim_{x\to 0^+} F(x) = 0$. See, e.g., https://ibmathsresources.com/2016/06/06/the-six-function/), for graphical illustration of the Si function.

Consider now a different approach, namely, using the Taylor series expansion of $\sin(x)/x$ about zero. From Maple, this is

$$\frac{\sin r}{r} = 1 - \frac{1}{6}r^2 + \frac{1}{120}r^4 - \frac{1}{5040}r^6 + O\left(r^8\right),$$

so ignoring higher-order terms,

$$F(x) \approx \int_0^x \left(1 - \frac{1}{6}r^2 + \frac{1}{120}r^4 - \frac{1}{5040}r^6\right) dr = x - \frac{1}{18}x^3 + \frac{1}{600}x^5 - \frac{1}{35\,280}x^7,$$

which is accurate for x near zero, and from which it appears safe to conclude that

$$\lim_{x\to 0^+} F(x) = 0. \tag{2.183}$$

Clearly, this expansion cannot be used to approximate $\lim_{x\to\infty} F(x)$. How to do this is considered next.

Example 2.54 *The integral*

$$S = \int_0^\infty \frac{\sin x}{x}\, dx = \frac{\pi}{2} \tag{2.184}$$

arises, for example, in the proof of the inversion theorem for continuous random variables (see, e.g., Paolella, Intermediate Probability, p. 31), which itself is of utmost importance in probability theory, and is heavily used in, e.g., quantitative risk management, advanced portfolio optimization, and option pricing.

We first wish to show that (2.184) converges. Recall from (2.56) that $\lim_{x\to 0} x^{-1} \sin x = 1$, *so that* $\int_0^1 x^{-1} \sin x \; dx$ *is well defined, and it suffices to consider* $\lim_{M\to\infty} \int_1^M x^{-1} \sin x \; dx$. *Integration by parts with* $u = x^{-1}$ *and* $dv = \sin x \; dx$ *gives*

$$\int_1^M x^{-1} \sin x \, dx = \; - x^{-1} \cos x \Big|_1^M - \int_1^M x^{-2} \cos x \, dx.$$

Clearly, $\lim_{M\to\infty} (\cos M)/M = 0$, *so the first term is unproblematic. For the integral, note that*

$$\left|\frac{\cos x}{x^2}\right| \le \frac{1}{x^2} \quad \textit{and} \quad \int_1^\infty \frac{dx}{x^2} < \infty,$$

so that S converges, having used the (improper integral) comparison test, (2.197).

Showing that the integral equals $\pi/2$ *is a standard calculation via contour integration (see, e.g., Bak and Newman, 1997, p. 134), though derivation without complex analysis is*

also possible. As in Hijab (1997, p. 197) and Jones (2001, p. 192), we begin by defining $F(x) = \int_0^x r^{-1} \sin r \, dr$ *for* $x > 0$. *Observe that, from FTC (ii, b) (2.146),* $F'(x) = x^{-1} \sin x$.

Integration by parts (2.158) of

$$I(b,s) := \int_0^b e^{-sx} \frac{\sin x}{x} dx$$

with $u = e^{-sx}$, $dv = x^{-1} \sin x \, dx$, $du = -se^{-sx} dx$ *and* $v = F(x)$ *gives, from (2.183),*

$$I(b,s) = e^{-sb} F(b) + s \int_0^b F(x) e^{-sx} dx = e^{-sb} F(b) + \int_0^{sb} F(y/s) e^{-y} \, dy,$$

with $y = sx$. *Letting* $b \to \infty$ *and using the boundedness of* $F(b)$ *as shown above gives*

$$I(b,s) \to I(\infty, s) = \int_0^\infty F(y/s) e^{-y} dy.$$

Now taking the limit in s, *and assuming the validity of the exchange of limit and integral,*

$$\lim_{s \to 0^+} I(\infty, s) = \int_0^\infty F(\infty) e^{-y} dy = F(\infty) \int_0^\infty e^{-y} dy = F(\infty),$$

from (2.170). But using (2.180) and the fact, proven in (2.271), that

$$\int_0^\infty e^{-sx} \frac{\sin x}{x} dx = \arctan\left(s^{-1}\right),$$

we have

$$\lim_{s \to 0^+} I(\infty, s) = \lim_{s \to 0^+} \int_0^\infty e^{-sx} \frac{\sin x}{x} dx = \lim_{s \to 0^+} \left(\arctan\left(s^{-1}\right)\right) = \arctan(\infty) = \frac{\pi}{2},$$

i.e.,

$$\frac{\pi}{2} = \lim_{s \to 0^+} I(\infty, s) = F(\infty) = \int_0^\infty \frac{\sin r}{r} dr,$$

as was to be shown. Other elementary proofs are outlined in Beardon (1997, p. 182) and Lang (1997, p. 343), while Goldberg (1964, p. 192) demonstrates that $\int_0^\infty \frac{\sin x}{x} dx$ *does not converge absolutely, i.e.,* $\int_0^\infty \frac{|\sin x|}{x} dx$ *does not exist. As an aside, it can also be shown that* $\int_0^\infty (\sin ax)/x \, dx = \operatorname{sgn}(a)\pi/2$. ■

Recall the tan and arctan functions, and their derivatives, e.g., (2.60) and (2.61). Before commencing to the next example, we show some basic results for the arctangent function. In particular, we will explicitly require knowing that $\arctan 1 = \pi/4$. Some analysis books (e.g., Ghorpade and Limaye, 2018, p. 246) *define* the arctangent function by

$$\arctan x := \int_0^x \frac{dt}{1 + t^2}. \tag{2.185}$$

The reason is that, from this, a rigorous, analytic (as opposed to geometric) definition of π can be elicited; and then from which all the usual trigonometric results can be rigorously derived, without any appeal to the area of a unit circle. From (2.185) and (2.139), we immediately see that

$$\arctan 0 = 0. \tag{2.186}$$

The rest of this presentation is based on Ghorpade and Limaye (2018).

As $1/(1+t^2) \geq 0$ for all $t \in \mathbb{R}$, $\arctan x \geq 0$ if $x > 0$, while $\arctan x \leq 0$ if $x < 0$. From FTC (ii,b) (2.146), the derivative of arctan is obviously positive on $\mathbb{R}$, so that arctan is strictly increasing on $\mathbb{R}$. Further,

$$(\arctan)''x = -\frac{2x}{(1+x^2)^2}$$

is positive for all $x \in (-\infty, 0)$ and negative for all $x \in (0, \infty)$. Thus, arctan is strictly convex on $(-\infty, 0)$ and strictly concave on $(0, \infty)$; and 0 is a point of inflection.

For $x \in \mathbb{R}$, and with $s = -t$,

$$\arctan(-x) = \int_0^{-x} \frac{1}{1+t^2} dt = -\int_0^x \frac{1}{1+s^2} ds = -\arctan x.$$

Hence, arctan is an odd function. For $x \in (1, \infty)$,

$$\arctan x = \int_0^1 \frac{1}{1+t^2} dt + \int_1^x \frac{1}{1+t^2} dt$$

and since $1 \geq t^2$ for $t \in [0, 1]$, while $t^2 \geq 1$ for $t \in [1, x]$, we see that

$$\int_0^1 \frac{1}{1+1} dt + \int_1^x \frac{1}{t^2+t^2} dt \leq \arctan x \leq \int_0^1 \frac{1}{1} dt + \int_1^x \frac{1}{t^2} dt.$$

The definite integrals above are easy to evaluate, and thus we obtain

$$1 - \frac{1}{2x} \leq \arctan x \leq 2 - \frac{1}{x} \quad \text{for all } x \in (1, \infty).$$

As arctan is strictly increasing and odd, $\forall x \in \mathbb{R}$, $-2 < \arctan x < 2$, i.e., arctan is a bounded function.

Function arctan is one-one because it is strictly increasing. We wish to show that it is also onto, and thus a bijection. Consider $y \in (-\pi/2, \pi/2)$. Since $\arctan x \to -\pi/2$ as $x \to -\infty$, there is some $x_0 \in \mathbb{R}$ such that $\arctan x_0 < y$, and since $\arctan x \to \pi/2$ as $x \to \infty$, there is some $x_1 \in \mathbb{R}$ such that $y < \arctan x_1$. From FTC (ii,a) (2.145), arctan is continuous on the interval $[x_0, x_1]$. Thus, the IVT (2.30) shows that $\exists x \in (x_0, x_1)$ such that $\arctan x = y$. Thus the function $\arctan : \mathbb{R} \to (-\pi/2, \pi/2)$ is bijective.

For $x \in [1, \infty)$, the substitution $t = 1/s$ gives

$$\int_1^x \frac{dt}{1+t^2} = \int_{1/x}^1 \frac{ds}{1+s^2}, \quad \text{and hence} \quad \lim_{x\to\infty} \int_1^x \frac{dt}{1+t^2} = \int_0^1 \frac{ds}{1+s^2},$$

so that

$$\lim_{x\to\infty} (\arctan x - \arctan 1) = \lim_{x\to\infty} \int_1^x \frac{dt}{1+t^2} = \arctan 1.$$

But $\arctan x \to \pi/2$ as $x \to \infty$, so that

$$\arctan 1 = \pi/4. \tag{2.187}$$

Example 2.55 *Example 2.49 showed that* $I = \int_0^\infty \exp\left(-x^2\right) dx$ *is convergent. Its value is commonly and quickly derived by use of polar coordinates (see §5.6, and Example 5.20), but it can be done without them. As in Weinstock (1990), let*

$$I = \int_0^\infty \exp\left(-x^2\right) dx$$

and

$$f(x) = \int_0^1 \frac{\exp\left(-x\left(1+t^2\right)\right)}{1+t^2} dt, \quad x > 0. \tag{2.188}$$

From (2.61) and FTC (i) (2.143); or from (2.185); and recalling (2.186) and (2.187), we have

$$f(0) = \int_0^1 \left(1+t^2\right)^{-1} dt = \arctan(1) - \arctan(0) = \pi/4 \tag{2.189}$$

and, as $x > 0$, *and* $0 < t < 1 \Rightarrow e^{-x\cdot 1} < e^{-xt} < e^{-x\cdot 0} = 1$,

$$0 < f(x) = e^{-x} \int_0^1 \frac{e^{-xt^2}}{1+t^2} dt < e^{-x} \int_0^1 \frac{1}{1+t^2} dt = \frac{\pi}{4} e^{-x},$$

so that, from the Squeeze Theorem, $f(\infty) = 0$. *Differentiating (2.188) with respect to* x *(and assuming we can interchange derivative and integral; see §5),* $f'(x)$ *is given by*

$$\int_0^1 \frac{d}{dx} \frac{\exp\left(-x\left(1+t^2\right)\right)}{1+t^2} dt = -\int_0^1 \exp\left(-x\left(1+t^2\right)\right) dt = -e^{-x} \int_0^1 \exp\left(-xt^2\right) dt.$$

Now, with $u = t\sqrt{x}$, $t = u/\sqrt{x}$ *and* $dt = x^{-1/2}\, du$,

$$f'(x) = -e^{-x} x^{-1/2} \int_0^{\sqrt{x}} \exp\left(-u^2\right) du = -e^{-x} x^{-1/2} g\left(\sqrt{x}\right), \tag{2.190}$$

where $g(z) := \int_0^z \exp\left(-u^2\right) du$. *From (2.189) and that* $f(\infty) = 0$, *integrating both sides of (2.190) from* 0 *to* ∞ *and using FTC (i) (2.143), gives, with* $z = \sqrt{x}$, $x = z^2$ *and* $dx = 2z\, dz$,

$$0 - \frac{\pi}{4} = f(\infty) - f(0) = -\int_0^\infty e^{-x} x^{-1/2} g\left(\sqrt{x}\right) dx = -2 \int_0^\infty e^{-z^2} g(z)\, dz$$

or $\int_0^\infty \exp(-z^2) g(z)\, dz = \pi/8$. *From FTC (ii,b) (2.146),* $g'(z) = \exp(-z^2)$, *so*

$$\frac{\pi}{8} = \int_0^\infty g'(z)\, g(z)\, dz \overset{u=g(z)}{=} \int_0^I u\, du = \frac{I^2}{2},$$

or $I = \sqrt{\pi}/2$.

It is now easy to show that $J = \int_0^\infty \exp\left(-u^2/2\right) du = \sqrt{\pi}/\sqrt{2}$: *With* $u = x\sqrt{2}$, $x = u/\sqrt{2}$, $dx = du/\sqrt{2}$,

$$\frac{\sqrt{\pi}}{2} = I = \int_0^\infty \exp\left(-x^2\right) dx = \frac{1}{\sqrt{2}} \int_0^\infty \exp\left(-u^2/2\right) du = \frac{1}{\sqrt{2}} \frac{\sqrt{\pi}}{\sqrt{2}}. \tag{2.191}$$

We will investigate two further ways of determining J *in Examples 5.20 and 5.21. Weinstock also gives similar derivations of the* Fresnel integrals $\int_0^\infty \cos y^2 dy$ *and* $\int_0^\infty \sin y^2 dy$. *Another way of calculating* $\int_{-\infty}^\infty \exp\left(-x^2\right) dx$ *without use of polar coordinates is detailed in Hijab (1997, §5.4).* ■

We return now to some theory. This first part of this material comes from Ghorpade and Limaye, pp. 91-2, and is included because it is anyway relevant, but also because it is needed for the proof of the Cauchy Criterion for improper integrals below, in (2.196).

Definition: Suppose $D \subseteq \mathbb{R}$ is such that D is not bounded above. Then there is a sequence in D that tends to ∞. Consider a function $f : D \to \mathbb{R}$. We say that a limit of f as x tends to infinity exists if there is a real number ℓ such that

$$(x_n) \text{ any sequence in } D \text{ and } x_n \to \infty \Longrightarrow f(x_n) \to \ell.$$

We then write

$$f(x) \to \ell \text{ as } x \to \infty \quad \text{or} \quad \lim_{x\to\infty} f(x) = \ell.$$

Since there exists a sequence (x_n) in D such that $x_n \to \infty$, it follows from the uniqueness of limits (see the theorem around (2.2)) that, if $\lim_{x\to\infty} f(x)$ exists, then it is unique.

Proposition: Suppose $D \subseteq \mathbb{R}$ is not bounded above and $f : D \to \mathbb{R}$ is a function. Then $\lim_{x\to\infty} f(x)$ exists if and only if there is $\ell \in \mathbb{R}$ satisfying the following $\epsilon - \alpha$ condition: For every $\epsilon > 0$, there is $\alpha \in \mathbb{R}$ such that

$$x \in D \text{ and } x \geq \alpha \Longrightarrow |f(x) - \ell| < \epsilon. \tag{2.192}$$

Proof: Assume that $\lim_{x\to\infty} f(x)$ exists and is equal to a real number ℓ. Suppose for a moment that the $\epsilon - \alpha$ condition does not hold. This means that there is $\epsilon > 0$ such that for every $\alpha \in \mathbb{R}$, there is $x \in D$ satisfying $x \geq \alpha$, but $|f(x) - \ell| \geq \epsilon$. By choosing $\alpha = n$ for each $n \in \mathbb{N}$, we may obtain a sequence (x_n) in D such that $x_n \geq n$, but $|f(x_n) - \ell| \geq \epsilon$ for all $n \in \mathbb{N}$. Now $x_n \to \infty$ and $f(x_n) \nrightarrow \ell$. This contradicts the assumption that $\lim_{x\to\infty} f(x) = \ell$.

Conversely, assume the $\epsilon - \alpha$ condition. Let (x_n) be any sequence in D such that $x_n \to \infty$. Let $\epsilon > 0$ be given. Then there is $\alpha \in \mathbb{R}$ such that

$$x \in D \text{ and } x \geq \alpha \Longrightarrow |f(x) - \ell| < \epsilon.$$

Since $x_n \to \infty$, there is $n_0 \in \mathbb{N}$ such that $x_n \geq \alpha$, and hence $|f(x_n) - \ell| < \epsilon$, for all $n \geq n_0$. Thus $f(x_n) \to \ell$. So $\lim_{x\to\infty} f(x)$ exists and equals ℓ.

Proposition (Cauchy Criterion): Suppose $D \subseteq \mathbb{R}$ is not bounded above and $f : D \to \mathbb{R}$ is a function. Then $\lim_{x\to\infty} f(x)$ exists if and only if the following $\epsilon - \alpha$ condition holds: For every $\epsilon > 0$, there is $\alpha \in \mathbb{R}$ such that

$$x, y \in D, \ \ x \geq \alpha, \text{ and } y \geq \alpha \Longrightarrow |f(x) - f(y)| < \epsilon. \tag{2.193}$$

Proof: Assume that $\lim_{x\to\infty} f(x)$ exists and is equal to a real number ℓ. Let $\epsilon > 0$ be given. Then there is $\alpha \in \mathbb{R}$ such that

$$x \in D \text{ and } x \geq \alpha \Longrightarrow |f(x) - \ell| < \frac{\epsilon}{2}.$$

Hence for $x, y \in D$ satisfying $x \geq \alpha$ and $y \geq \alpha$, we obtain

$$|f(x) - f(y)| \leq |f(x) - \ell| + |\ell - f(y)| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Conversely, assume that the $\epsilon - \alpha$ condition holds. Let $\epsilon > 0$ be given. Then there is $\alpha \in \mathbb{R}$ such that

$$x, y \in D, \;\; x \geq \alpha \text{ and } y \geq \alpha \Longrightarrow |f(x) - f(y)| < \frac{\epsilon}{2}.$$

By our hypothesis, there is a sequence (x_n) in D such that $x_n \to \infty$. Hence there is $n_0 \in \mathbb{N}$ such that $x_n \geq \alpha$ for all $n \geq n_0$. Consequently,

$$\forall n, m \geq n_0, \quad |f(x_n) - f(x_m)| < \frac{\epsilon}{2}.$$

Thus $(f(x_n))$ is a Cauchy sequence in $\mathbb{R}$. By the Cauchy Criterion for sequences (2.198), there is $\ell \in \mathbb{R}$ such that $f(x_n) \to \ell$. Hence there is $n_1 \in \mathbb{N}$ such that $n_1 \geq n_0$ and $|f(x_{n_1}) - \ell| < \epsilon/2$. Since $x_{n_1} \geq \alpha$, it follows that

$$x \in D \text{ and } x \geq \alpha \Longrightarrow |f(x) - \ell| \leq |f(x) - f(x_{n_1})| + |f(x_{n_1}) - \ell| < \epsilon.$$

Consequently, by Proposition (2.192), $\lim_{x\to\infty} f(x)$ exists and is equal to ℓ.

The following is based on Ghorpade and Limaye, §9.4.

Let $a \in \mathbb{R}$, and let f be defined on $[a, \infty)$ and such that f is integrable on $[a, x]$ for every $x \geq a$. It is useful to define

$$F(x) = \int_a^x f(t)dt \quad \text{for all } x \in [a, \infty),$$

with F called the partial integral function, and $F(x)$ the partial integral of f.

Proposition: If $\int_{t\geq a} f(t)dt$ is convergent, then the set $\left\{\int_a^x f(t)dt : x \in [a, \infty)\right\}$ of partial integrals is bounded.

Proof: Let $F(x) := \int_a^x f(t)dt$ for $x \in [a, \infty)$ and note that since $\int_{t\geq a} f(t)dt$ is convergent, there exists $x_0 \geq a$ such that

$$\forall x \geq x_0, \quad |F(x)| \leq 1 + \left|\int_a^\infty f(t)dt\right|. \tag{2.194}$$

We add a bit to the previous proof. Convergence of $\int_a^\infty f(t)dt$ means, $\exists L \in \mathbb{R}$ such that $\lim_{x\to\infty} \int_a^x f(t)dt = L$, i.e., for any given $\epsilon > 0$, $\exists x_0 \geq a$ such that, for all $x \geq x_0$,

$$\left|L - \int_a^x f(t)dt\right| < \epsilon,$$

so that $F(x)$ is bounded for all $x \geq x_0$. By definition, f is integrable on $[a, x]$ for every $x \geq a$, and from the definition of the Riemann integral in §2.4.1, $F(x)$ is finite for all $a \leq x < x_0$. Recall the reverse triangle inequality (1.8), namely, $\forall a, b \in \mathbb{R}$, $||a| - |b|| \leq |a + b|$ or $\|a| - |b|| \leq |b - a|$. Then

$$\big|\; |F(x)| - |L| \;\big| = \left|\; \left|\int_a^x f(t)dt\right| - |L| \;\right| \leq \left|L - \int_a^x f(t)dt\right| < \epsilon, \tag{2.195}$$

i.e., $-\epsilon < |F(x)| - |L| < \epsilon$, or $|F(x)| < \epsilon + |L|$, which is (2.194).

Proposition (Cauchy Criterion): An improper integral $\int_{t\geq a} f(t)dt$ is convergent if and only if for every $\epsilon > 0$, there exists $x_0 \in [a, \infty)$ such that

$$\left| \int_x^y f(t)dt \right| < \epsilon \quad \text{for all } y \geq x \geq x_0. \tag{2.196}$$

Proof: As in Ghorpade and Limaye, p. 394, let F denote the partial integral function of f, and write $F(y) - F(x) = \int_x^y f(t)dt$ for all $y \geq x \geq x_0$. Now use the Cauchy Criterion for unbounded sets, (2.193).

Proposition (Comparison Test for Improper Integrals): Suppose $a \in \mathbb{R}$ and $f, g : [a, \infty) \to \mathbb{R}$ are such that both f and g are integrable on $[a, x]$ for every $x \geq a$ and $|f(t)| \leq g(t)$ for all large $t \in [a, \infty)$.

$$\textbf{If } \int_{t\geq a} g(t)dt \textbf{ is convergent, then } \int_{t\geq a} f(t)dt \textbf{ is absolutely convergent.} \tag{2.197}$$

Proof: Let $t_0 \in [a, \infty)$ be such that $|f(t)| \leq g(t)$ for all $t \in [t_0, \infty)$. Suppose $\int_{t\geq a} g(t)dt$ is convergent. Let $\epsilon > 0$ be given. Then by the Cauchy Criterion, there exists $x_0 \in [a, \infty)$ such that $\left|\int_x^y g(t)dt\right| < \epsilon$ for all $y \geq x \geq x_0$. Now let $x_1 := \max\{t_0, x_0\}$. Then

$$\forall y \geq x \geq x_1, \quad \int_x^y |f(t)|dt \leq \int_x^y g(t)dt = \left| \int_x^y g(t)dt \right| < \epsilon.$$

Hence by the Cauchy Criterion (2.196), $\int_{t\geq a} f(t)dt$ is absolutely convergent.

As a simple example, to prove the existence of the integral $\int_1^\infty \frac{dx}{1+x^2}$ we simply have to observe that, for any $x \geqslant 1$, $\frac{1}{1+x^2} \leqslant \frac{1}{x^2}$. Since $\int_1^\infty x^{-2}dx$ exists, the result then follows from (2.197).

Proposition (Limit Comparison Test for Improper Integrals): Let $a \in \mathbb{R}$ and let $f, g : [a, \infty) \to \mathbb{R}$ be such that both f and g are integrable on $[a, x]$ for every $x \geq a$. Suppose $f(t) > 0$ and $g(t) > 0$ for all large $t \in [a, \infty)$. Also suppose there exists $\ell \in \mathbb{R} \cup \{\infty\}$ such that $f(t)/g(t) \to \ell$ as $t \to \infty$.

(i) If $\ell \neq 0$ and $\ell \neq \infty$, then

$$\int_{t\geq a} f(t)dt \text{ is convergent } \iff \int_{t\geq a} g(t)dt \text{ is convergent.}$$

(ii) If $\ell = 0$ and $\int_{t\geq a} g(t)dt$ is convergent, then $\int_{t\geq a} f(t)dt$ is convergent.

(iii) If $\ell = \infty$ and $\int_{t\geq a} f(t)dt$ is convergent, then $\int_{t\geq a} g(t)dt$ is convergent.

A proof can be found in Ghorpade and Limaye, p. 402.

Example 2.56 *Let $q \in \mathbb{R}$ and let $f : [1, \infty) \to \mathbb{R}$ be given by $f(t) := e^{-t}t^q$. Then $\int_{t\geq 1} f(t)dt$ is convergent. To see this, choose $k \in \mathbb{N}$ with $k > q + 1$, and define $g : [1, \infty) \to \mathbb{R}$ by $g(t) := t^{q-k}$. Then $f(t) > 0$ and $g(t) > 0$ for all $t \in [1, \infty)$, and moreover, by L'Hôpital's Rule,*

$$\frac{f(t)}{g(t)} = \frac{t^k}{e^t} \to 0 \quad \text{as} \quad t \to \infty.$$

Since $k - q > 1$, we see that $\int_{t\geq 1} g(t)dt$ is convergent. Hence by part (ii) of the previous proposition, we conclude that $\int_{t\geq 1} e^{-t}t^q dt$ is convergent. ■

2.5 Series

Of particular importance to us is differentiation and integration of infinite series of functions, along with the development of Taylor polynomials and Taylor series of functions. Their development, however, first requires establishing various notions and tools, of interest in themselves, such as sequences, Cauchy sequences, series of numbers, Cauchy products, and the highly important notions of lim inf and lim sup.

2.5.1 Definitions and Examples

> Mathematics is an experimental science, and definitions do not come first, but later on. (Oliver Heaviside)

Recall the beginning of §2.1, in which we outlined the basic definitions and results regarding sequences of numbers. We build upon that material now. We begin with Cauchy sequences.

Definition: Sequence $\{s_n\}$ is termed a *Cauchy sequence* if, for a given $\epsilon \in \mathbb{R}_{>0}$, $\exists\, N \in \mathbb{N}$ such that $\forall n, m \geq N$, $|s_m - s_n| < \epsilon$.

Theorem:

$$\textbf{A sequence } \{s_n\} \textbf{ converges} \iff \{s_n\} \textbf{ is a Cauchy sequence.} \tag{2.198}$$

This is a fundamental result, proven in all beginning real analysis books.

Recall from §1.1 the definitions of infimum and supremum; which we repeat here: Let S be a nonempty subset of $\mathbb{R}$. We say S has an *upper bound* M if $x \leq M\ \forall x \in S$, in which case S is *bounded above* by M. Note that, if S is bounded above, then it has infinitely many upper bounds. A fundamental property of $\mathbb{R}$ not shared by $\mathbb{Q}$ is that, if S is a nonempty set that has an upper bound M, then S possesses a unique *least upper bound*, or *supremum*, denoted $\sup S$. That is, $\exists\, U \in \mathbb{R}$ such that U is an upper bound of S, and such that, if V is also an upper bound of S, then $V \geq U$. If S is not bounded above, then $\sup S = \infty$. Also, $\sup \emptyset = -\infty$. Similar terminology applies to the *greatest lower bound*, or *infimum* of S, denoted $\inf S$.

Theorem: If $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ are two sequences of real numbers that are each bounded above, then

$$\sup_{n\in\mathbb{N}} \{a_n + b_n\} \leq \sup_{n\in\mathbb{N}} a_n + \sup_{n\in\mathbb{N}} b_n. \tag{2.199}$$

> Proof: For simplicity of notation, set $u = \sup a_n$ and $v = \sup b_n$. Since $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ are each bounded above, we know that u and v are finite real numbers. Since u is an upper bound for the a_n, we know that $a_n \leq u$ for every n. Similarly, $b_n \leq v$ for every n. Therefore $a_n + b_n \leq u + v$ for every n. Hence $u + v$ is an upper bound for $(a_n + b_n)_{n\in\mathbb{N}}$, so this sequence is bounded above and therefore has a finite supremum, which we will denote by $w = \sup(a_n + b_n)$. By definition, w is the least upper bound for the sequence $(a_n + b_n)_{n\in\mathbb{N}}$. Since we saw above that $u + v$ is an upper bound for $(a_n + b_n)_{n\in\mathbb{N}}$, we must therefore have $w \leq u + v$, which is exactly what we wanted to prove.

Theorem: Let $\{a_n\}, \{b_n\} \in \mathbb{R}_{\geq 0}$, for $n \in \mathbb{N}$. Then

$$\sup_{n\in\mathbb{N}} a_n b_n \leq \sup_{n\in\mathbb{N}} a_n \sup_{n\in\mathbb{N}} b_n. \tag{2.200}$$

and, more generally, for each $n \in \mathbb{N}$,

$$\sup_{k\geq n} a_k b_k \leq \sup_{k\geq n} a_k \sup_{k\geq n} b_k. \tag{2.201}$$

Proof: Clearly, (2.200) follows from (2.201), so we only need to prove the latter. For each $n \in \mathbb{N}$, $a_n \leq \sup_{k\geq n} a_k$ and $b_n \leq \sup_{k\geq n} b_k$, so that, as $\{a_n\}, \{b_n\} \in \mathbb{R}_{\geq 0}$, $a_n b_n \leq \left(\sup_{k\geq n} a_k\right)\left(\sup_{k\geq n} b_k\right)$. As the rhs is an upper bound, (2.201) follows.

Theorem: Let A be a nonempty subset of $\mathbb{R}_{\geq 0}$, i.e., a set (countable or uncountable) of nonnegative real numbers. Then

$$\sup_{a\in A} a^2 = \left(\sup_{a\in A} a\right)^2, \tag{2.202}$$

i.e., "the sup of the square equals the square of the sup."

Proof: If $\sup\{a\} = \infty$, then (2.202) is immediate. Assume $\sup\{a\} \in \mathbb{R}$. We need to demonstrate that (i) $\sup\{a^2\} \leq (\sup\{a\})^2$ and (ii) $(\sup\{a\})^2 \leq \sup\{a^2\}$. We suppress writing $a \in A$, i.e., we just write $\{a^2\}$ to mean $\{a^2\}_{a\in A}$. Similarly, we suppress always writing $\forall a \in A$, i.e., $\sup\{a\} \geq a$ is short for $\forall a \in A$, $\sup\{a\} \geq a$.

For (i): Note that $\sup\{a\} \geq a$ (for all $a \in A$), hence $(\sup\{a\})^2 \geq a^2$ (for all $a \in A$), so $(\sup\{a\})^2$ is an upper bound for $\{a^2\} = \{a^2\}_{a\in A}$, hence it is greater than the least upper bound, i.e., $(\sup\{a\})^2 \geq \sup\{a^2\}$.

For (ii): Let N be large enough such that $\sup\{a\} - \frac{1}{N} > 0$ (if $\sup\{a\} = 0$ and all a are non-negative, then it is clear what the set is and the result itself). Now for any $n > N$, the quantity $\sup\{a\} - \frac{1}{n}$ is not an upper bound of $\{a\}$, Hence, there is some a_n such that $a_n > \sup\{a\} - \frac{1}{n}$. Square both sides(note that by non-negativity of both sides, this preserves sign) to see that $a_n^2 > \frac{1}{n^2} + (\sup\{a\})^2 - \frac{2\sup\{a\}}{n}$. Now, by definition of supremum, we have

$$\sup\{a^2\} \geq a_n^2 > \frac{1}{n^2} + (\sup\{a\})^2 - \frac{2\sup\{a\}}{n}.$$

This applies for all $n > N$. Since $(\sup\{a\})^2$ is a bounded quantity, letting $n \to \infty$, we see that $\sup\{a^2\} \geq (\sup\{a\})^2$. Hence, equality follows.

We will use (2.202) to prove an important result below, namely (2.248).

Let $\{s_n\}$ be a sequence. For each $k \in \mathbb{N}$, define the two sequences $a_k = \inf\{s_n : n \geq k\}$ and $b_k = \sup\{s_n : n \geq k\}$. A bit of thought confirms that a_k is increasing and b_k is decreasing. Thus, if they are bounded, then they converge to a value in $\mathbb{R}$, and if they are not bounded, then they diverge to plus or minus ∞. Either way, the two sequences have limits in $\mathbb{X}$. The *limit supremum* (or *limit superior*) of $\{s_n\}$, denoted $\limsup s_n$, is

$$\limsup_{n\to\infty} s_n = \lim_{k\to\infty} b_k = \lim_{k\to\infty} \left(\sup_{n\geq k} s_k\right),$$

and the *limit infimum* (or *limit inferior*), denoted $\liminf s_n$, is

$$\liminf_{n\to\infty} s_n = \lim_{k\to\infty} a_k = \lim_{k\to\infty} \left(\inf_{n\geq k} s_k\right).$$

Because a_k is increasing, and b_k is decreasing, it follows that

$$\liminf_{n\to\infty} s_n = \sup_{k\in\mathbb{N}} \inf_{n\geq k} s_k \quad \text{and} \quad \limsup_{n\to\infty} s_n = \inf_{k\in\mathbb{N}} \sup_{n\geq k} s_k.$$

The following facts are straightforward to prove, and should at least be informally thought through until they become intuitive:

- $\limsup s_n = -\infty$ iff $\lim_{n\to\infty} s_n = -\infty$.
- $\limsup s_n = \infty$ iff, $\forall M \in \mathbb{R}$ and $n \in \mathbb{N}$, $\exists k \in \mathbb{N}$ with $k \geq n$ such that $s_k \geq M$.[14]
- Suppose $\limsup s_n \in \mathbb{R}$ (i.e., it is finite). Then, $\forall \epsilon > 0$, $U = \limsup s_n$ iff

$$\exists N \in \mathbb{N} \text{ such that, } \forall n \geq N,\ s_n < U + \epsilon, \tag{2.203}$$

and

$$\text{Given } n \in \mathbb{N},\ \exists k \in \mathbb{N} \text{ with } k \geq n \text{ such that } s_k > U - \epsilon. \tag{2.204}$$

(to verify these, think what would happen if they were not true.)

Similarly,

- $\liminf s_n = \infty$ iff $\lim_{n\to\infty} s_n = \infty$.
- $\liminf s_n = -\infty$ iff, $\forall M \in \mathbb{R}$ and $n \in \mathbb{N}$, $\exists k \in \mathbb{N}$ with $k \geq n$ such that $s_k \leq M$.
- Suppose $\liminf s_n \in \mathbb{R}$. Then, given any $\epsilon > 0$, $L = \liminf s_n$ iff

$$\exists N \in \mathbb{N} \text{ such that, } \forall n \geq N,\ s_n > L - \epsilon,$$

and

$$\text{Given } n \in \mathbb{N},\ \exists k \in \mathbb{N} \text{ with } k \geq n \text{ such that } s_k < L + \epsilon.$$

It is easy to see from the definitions that

$$\textbf{If } \ell = \lim_{n\to\infty} s_n \textbf{ for } \ell \in \mathbb{X}\textbf{, then } \liminf s_n = \limsup s_n = \ell\textbf{.} \tag{2.205}$$

The converse of (2.205) is also true:

$$\textbf{If } \liminf s_n = \limsup s_n = \ell \textbf{ for } \ell \in \mathbb{X}\textbf{, then } \ell = \lim_{n\to\infty} s_n\textbf{.} \tag{2.206}$$

To prove (2.206) for $\ell \in \mathbb{R}$, note that, for a given $\epsilon > 0$, the above results imply that $\exists\, n_1, n_2 \in \mathbb{N}$ such that, $\forall n \geq n_1$, $s_n < \ell + \epsilon$ and, $\forall n \geq n_2$, $s_n > \ell - \epsilon$. Thus, with $N = \max\{n_1, n_2\}$, $\ell - \epsilon < s_n < \ell + \epsilon$ for all $n \geq N$, or $\lim_{n\to\infty} s_n = \ell$.

Theorem: Let $\{a_n\}, \{b_n\} \in \mathbb{R}_{\geq 0}$. Then

$$\limsup\{a_n b_n\} \leq \limsup\{a_n\} \limsup\{b_n\}. \tag{2.207}$$

Proof: Recall (2.201), i.e., $\sup_{k\geq n} a_k b_k \leq \sup_{k\geq n} a_k \sup_{k\geq n} b_k$. Let $c_k = a_k b_k$ and $\bar{c}_n = \sup_{k\geq n} c_k$, $\bar{a}_n = \sup_{k\geq n} a_k$, and $\bar{b}_n = \sup_{k\geq n} b_k$. Then (2.201) reads, for sequences $\{\bar{c}_n\}$, $\{\bar{a}_n\}$, and $\{\bar{b}_n\}$,

$$\forall n \in \mathbb{N}, \quad \bar{c}_n \leq \bar{a}_n \bar{b}_n. \tag{2.208}$$

[14] Note the difference to saying that $\lim_{n\to\infty} s_n = \infty$: If $s_n \to \infty$, then $\limsup s_n = \infty$, but the converse need not be true.

Then, from (2.208); (2.7) (which says, for x_n and y_n sequences such that $\lim_{n\to\infty} x_n = x$ and $\lim_{n\to\infty} y_n = y$, if $x_n \le y_n$ for all n sufficiently large, then $x \le y$); and that the limit of a product is the product of the limits,

$$\begin{aligned}\limsup\{a_n b_n\} = \lim_{n\to\infty} \sup_{k\ge n} a_k b_k = \lim_{n\to\infty} \bar{c}_n \le \lim_{n\to\infty} \bar{a}_n \bar{b}_n \\ = \lim_{n\to\infty} \sup_{k\ge n} a_k \sup_{k\ge n} b_k = \limsup\{a_n\} \limsup\{b_n\},\end{aligned}$$

which is (2.207).

Relation (2.207) can be strict. As an example, take $a_n = ((-1)^{n+1} + 1)/2$, i.e., $\{a_n\} = \{1, 0, 1, 0, \ldots\}$; and $b_n = ((-1)^n + 1)/2$, i.e., $\{b_n\} = \{0, 1, 0, 1, \ldots\}$. This results in a strict inequality.

Definition: (Series; and Convergent Series, Divergent Series). A *series* is an infinite sum of elements from a set, such as the real numbers, the complex numbers, vectors, matrices, functions, etc.. We say that a series

$$\sum_{n=1}^{\infty} c_n = c_1 + c_2 + \cdots$$

of real numbers *converges* if there is a real number s such that the partial sums

$$s_N = \sum_{n=1}^{N} c_n = c_1 + c_2 + \cdots + c_N$$

converge to s as $N \to \infty$. In this case we declare that the series $\sum_{n=1}^{\infty} c_n$ has the value s:

$$\sum_{n=1}^{\infty} c_n = \lim_{N\to\infty} s_N = \lim_{N\to\infty} \sum_{n=1}^{N} c_n = s.$$

If the series $\sum_{n=1}^{\infty} c_n$ does not converge, then we say that it *diverges*.

Theorem (The nth Term Test): If $\sum_{n=1}^{\infty} c_n$ is a convergent series of real numbers, then

$$\lim_{n\to\infty} c_n = 0. \tag{2.209}$$

Proof: Since the series converges, $x = \sum_{n=1}^{\infty} c_n$ is a real number. Let $s_N = \sum_{n=1}^{N} c_n$ be the Nth partial sum of the series, and set $s_0 = 0$. Then, since s_n converges to x as $n \to \infty$, we have, using the linearity property of limits (2.12),

$$\lim_{n\to\infty} c_n = \lim_{n\to\infty} (s_n - s_{n-1}) = \lim_{n\to\infty} s_n - \lim_{n\to\infty} s_{n-1} = x - x = 0.$$

The converse of (2.209) is not true (use $\sum_{k=1}^{\infty} k^{-1}$ as an example).

Theorem (Tails of Convergent Series): If $\sum_{n=1}^{\infty} c_n$ is a convergent series of real numbers, then

$$\lim_{N\to\infty} \left(\sum_{n=N}^{\infty} c_n \right) = 0. \tag{2.210}$$

Proof: Let $x = \sum_{n=1}^{\infty} c_n$, and for each M let $s_M = \sum_{n=1}^{M} c_n$ be the M th partial sum of this series. Since the series converges, we know that $\lim_{M\to\infty} s_M = x$. Now let N be a fixed positive integer, and for each $M \geq N$ let

$$t_M = \sum_{n=N}^{M} c_n = \sum_{n=1}^{M} c_n - \sum_{n=1}^{N-1} c_n = s_M - s_{N-1}$$

be the M th partial sum of the infinite series $\sum_{n=N}^{\infty} c_n$. Keeping N fixed, we have that

$$\sum_{n=N}^{\infty} c_n = \lim_{M\to\infty} t_M = \lim_{M\to\infty} (s_M - s_{N-1}) = x - s_{N-1}.$$

Therefore

$$\lim_{N\to\infty} \sum_{n=N}^{\infty} c_n = \lim_{N\to\infty} (x - s_{N-1}) = x - x = 0.$$

Definition: Let f_k be a sequence of (real) functions with common domain. Let

$$s_n = \sum_{k=1}^{n} f_k, \quad \text{and} \quad S = \sum_{k=1}^{\infty} f_k = \lim_{n\to\infty} \sum_{k=1}^{n} f_k = \lim_{n\to\infty} s_n.$$

S is referred to as a *series* associated with the sequence $\{f_k\}$, and s_n is its nth partial sum. Series S *converges* if $\lim_{n\to\infty} s_n$ exists, i.e., if the limit is bounded, and *diverges* if the partial sums are not bounded.

Theorem: Let f_k be a sequence of (real) functions with common domain D.

$$\textbf{If series } S = \sum_{k=1}^{\infty} f_k \textbf{ converges, then } \lim_{n\to\infty} f_n = 0. \tag{2.211}$$

This follows by applying the nth term test (2.209) to $f(x)$ for each $x \in D$.

Here, 0 refers to the zero-function, i.e., $z(D) = 0$. The converse of (2.211), similar to series of real numbers, is not true.

If $\sum_{k=1}^{\infty} |f_k|$ converges, then S is said to *converge absolutely*. If S is absolutely convergent, then S is convergent, but the converse is not true. For example, the *alternating (harmonic) series* $S = \sum_{k=1}^{\infty} (-1)^k k^{-1}$ is convergent, but not absolutely convergent, i.e., as is proven in all real analysis books, the harmonic series

$$S = \sum_{k=1}^{\infty} k^{-1} \quad \text{diverges.} \tag{2.212}$$

It is *conditionally convergent*. The *Cauchy criterion* (Cauchy, 1821) is:

$$\sum_{k=1}^{\infty} f_k \textbf{ converges } \Leftrightarrow \exists N \in \mathbb{N} \textbf{ such that, } \forall n, m \geq N, \left| \sum_{k=n+1}^{m} f_k \right| < \epsilon, \tag{2.213}$$

for any given $\epsilon > 0$. This follows from (2.198) and writing $\left|\sum_{k=n+1}^{m} f_k\right| = |s_m - s_n|$.

The *geometric series* (with two common forms), and the *p-series* or *(Riemann's) zeta function*

$$S_0(c) = \sum_{k=0}^{\infty} c^k, \quad S_1 = \sum_{k=1}^{\infty} c^k, \ c \in [0,1), \quad \text{and} \quad \zeta(p) = \sum_{k=1}^{\infty} \frac{1}{k^p}, \ p \in \mathbb{R}_{>1},$$

respectively, are important convergent series because they can often be used to help prove the convergence of other series via the tests outlined below. Indeed, for the geometric series $S_1 = S_1(c) = c + c^2 + c^3 + \cdots$, $cS_1 = c^2 + c^3 + \cdots$ and $S_1 - cS_1 = c - \lim_{k\to\infty} c^k = c$, for $c \in [0,1)$. Solving $S_1 - cS_1 = c$ implies

$$S_1 = \frac{c}{1-c}, \quad c \in [0,1). \tag{2.214}$$

Trivially,

$$S_0 = 1 + \frac{c}{1-c} = \frac{1}{1-c}, \quad c \in [0,1). \tag{2.215}$$

Further, for $-1 < c \leq 0$,

$$S_0(c) = 1 - c + c^2 - c^3 + \cdots = \frac{1}{1-(-c)} = \frac{1}{1+c}. \tag{2.216}$$

Example 2.57 *For the zeta function, use (2.215) to express an upper bound of it as*

$$\begin{aligned} \zeta(p) = \sum_{k=1}^{\infty} \frac{1}{k^p} &= 1 + \frac{1}{2^p} + \frac{1}{3^p} + \cdots \\ &= 1 + \left(\frac{1}{2^p} + \frac{1}{3^p}\right) + \left(\frac{1}{4^p} + \frac{1}{5^p} + \frac{1}{6^p} + \frac{1}{7^p}\right) + \cdots \\ &< 1 + \frac{2}{2^p} + \frac{4}{4^p} + \cdots = \sum_{i=0}^{\infty} \left(\frac{1}{2^{p-1}}\right)^i = \frac{1}{1 - \frac{1}{2^{p-1}}}. \end{aligned}$$

This is valid for $1/2^{p-1} < 1$ or $(p-1)\ln 2 > 0$ or $p > 1$. Thus, we can conclude that the zeta function converges for at least $p > 1$, but we know from (2.212) that it diverges for $p = 1$, and thus the zeta function converges for $p > 1$ and (by use of the comparison test for series (2.224)) diverges for $0 \leq p \leq 1$. ■

Example 2.58 *Let $\zeta(p) = \sum_{r=1}^{\infty} r^{-p}$. The well-known result*

$$\zeta(2) = \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} \tag{2.217}$$

is often proven via contour integration (Bak and Newman, 1997, p. 141) or in the context of Fourier analysis (Stoll, 2001, p. 413; Jones, 2001, p. 410). The first proof, by Euler in 1735, involves the use of infinite products; see Havil (2003, p. 39) for a simple account, or Hijab (1997, §5.6) and Bak and Newman (1997, p. 223) for a rigorous proof. Before Euler solved it, the problem had been unsuccessfully attempted by Wallis, Leibniz, Jakob Bernoulli, and others (Havil, 2003, p. 38). Today, there are many known methods of proof.[15] *It can also be shown that $\zeta(4) = \pi^4/90$ and $\zeta(6) = \pi^6/945$. In general, expressions exist for even p.* ■

[15] Matsuoka (1961) gives an elementary one, requiring two integration by parts of $\int_0^{\pi/2} \cos^{2n}(t)\,dt = (\pi/2)(2n-1)!!\,/\,(2n)!!$, where $(2n)!! = 2 \cdot 4 \cdots (2n)$, $0!! = 1$, $(2n+1)!! = 1 \cdot 3 \cdots (2n+1)$ and $(-1)!! = 1$. Kortam (1996) illustrates further simple proofs, Hofbauer (2002) and Harper (2003) each contribute yet another method, and Chapman (2003) provides 14 proofs (not including the previous two, but including that from Matsuoka).

Example 2.59 *Recall* $e^\lambda = \lim_{n\to\infty}(1+\lambda/n)^n$ *from (2.104). From (2.81),* $e^\lambda = [\exp(1)]^\lambda$, *so that, to show convergence of the latter limit, it suffices to take* $\lambda = 1$ *and show that sequence* $s_n := (1+1/n)^n$ *converges. Applying the binomial theorem (1.21) to* s_n *gives* $s_n = \sum_{i=0}^n \binom{n}{i}(\frac{1}{n})^i$, *with each term expressible as*

$$\begin{aligned}\binom{n}{i}n^{-i} &= \frac{n_{[i]}}{i!}n^{-i} = \frac{n\,(n-1)\cdots(n-i+1)}{i!}n^{-i} \\ &= \frac{1}{i!}(1)\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{i-1}{n}\right). \end{aligned} \tag{2.218}$$

Similarly, $s_{n+1} = \sum_{i=0}^{n+1}\binom{n+1}{i}(n+1)^{-i}$, *with*

$$\begin{aligned}\binom{n+1}{i}(n+1)^{-i} &= \frac{(n+1)\,n\,(n-1)\cdots(n-i+2)}{i!}(n+1)^{-i} \\ &= \frac{1}{i!}(1)\left(\frac{n}{n+1}\right)\left(\frac{n-1}{n+1}\right)\cdots\left(\frac{n-i+2}{n+1}\right) \\ &= \frac{1}{i!}\left(1-\frac{1}{n+1}\right)\left(1-\frac{2}{n+1}\right)\cdots\left(1-\frac{i-1}{n+1}\right). \end{aligned} \tag{2.219}$$

As the quantity in (2.219) is larger than that in (2.218), it follows that $s_n \le s_{n+1}$, *i.e.,* s_n *is an increasing (or, nondecreasing) sequence. Also, for* $n \ge 2$,

$$s_n = \sum_{i=0}^n \binom{n}{i}\frac{1}{n^i} = \sum_{i=0}^n \frac{n\,(n-1)\cdots(n-i+1)}{n^i}\frac{1}{i!} \le \sum_{i=0}^n \frac{1}{i!}.$$

Note that $2! < 2^2$ *and* $3! < 2^3$, *but* $4! > 2^4$. *Assume* $k! > 2^k$ *holds for* $k \ge 4$. *It holds for* $k+1$ *because* $(k+1)! = (k+1)\,k! > (k+1)\,2^k > 2\cdot 2^k = 2^{k+1}$. *Thus,* $k! > 2^k$ *for* $k \ge 4$, *and*

$$\begin{aligned}\sum_{i=0}^n \frac{1}{i!} = \frac{8}{3} + \sum_{i=4}^n \frac{1}{i!} &< \frac{8}{3} + \sum_{i=4}^n \frac{1}{2^i} = \frac{8}{3} + \sum_{i=0}^n \frac{1}{2^i} - \sum_{j=0}^3 \frac{1}{2^j} = \frac{19}{24} + \sum_{i=0}^n \frac{1}{2^i} \\ &< \frac{19}{24} + \sum_{i=0}^\infty \frac{1}{2^i} = \frac{19}{24} + 2 < 2.8. \end{aligned}$$

Thus, s_n *is a nondecreasing, bounded sequence, and is thus convergent.* ■

We close this subsection by discussing the lim inf and lim sup for sets. We give two definitions, and show their equivalence. This material is not difficult, and will be essential when learning measure theory, the Lebesgue integral, and probability theory. In the following, we do not need to define measurable spaces or measure spaces, so just ignore this terminology for now.

Definition: Let $\{E_k\}_{k=1}^\infty$ be a countable family of measurable subsets of measurable space $(X, \mathcal{A})$. We define

$$\liminf_{k\to\infty} E_k := \bigcup_{n=1}^\infty \bigcap_{k=n}^\infty E_k, \quad \limsup_{k\to\infty} E_k := \bigcap_{n=1}^\infty \bigcup_{k=n}^\infty E_k. \tag{2.220}$$

Definition: Equivalent definitions are given by

$$\begin{aligned} \liminf_{k\to\infty} E_k &:= \big\{x \in X : x \in E_k \text{ for all but finitely many } k\big\}, \\ \limsup_{k\to\infty} E_k &:= \big\{x \in X : x \in E_k \text{ for infinitely many } k\big\} \end{aligned} \tag{2.221}$$

For $\liminf E_k$, "for all but finitely many k" means, $\exists k_0 \in \mathbb{N}$ such that, $\forall k \geq k_0$, $x \in E_k$. From (2.221), it is apparent that $\liminf E_k \subset \limsup E_k$.

Theorem: The two formulations (2.220) and (2.221) are equivalent.

Proof: We begin with ($\Rightarrow$) and ($\Leftarrow$) for lim sup.
($\Rightarrow$) Let $x \in \limsup$ in (2.220), so that

$$x \in \left(\bigcup_{k=1}^{\infty} E_k\right) \cap \left(\bigcup_{k=2}^{\infty} E_k\right) \cap \left(\bigcup_{k=3}^{\infty} E_k\right) \cap \cdots . \tag{2.222}$$

Suppose $x \notin \big\{x \in X : x \in E_k \text{ for infinitely many } k\big\}$. Then $\exists k_0 \in \mathbb{N}$ such that, $\forall k \geq k_0$, $x \notin E_k$, which implies $x \notin \bigcup_{k=k_0}^{\infty} E_k$, which contradicts (2.222).

($\Leftarrow$) Let $x \in \limsup$ in (2.221). Then x never stops reappearing in $\{E_k\} \Leftrightarrow x \in \bigcup_{k=1}^{\infty} E_k$, $x \in \bigcup_{k=2}^{\infty} E_k$, etc. $\Leftrightarrow$

$$x \in \left(\bigcup_{k=1}^{\infty} E_k\right) \cap \left(\bigcup_{k=2}^{\infty} E_k\right) \cap \cdots \Leftrightarrow x \in \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} E_k.$$

Now we do ($\Rightarrow$) and ($\Leftarrow$) for lim inf.
($\Rightarrow$) Let $x \in \liminf$ in (2.220), so that

$$x \in \left(\bigcap_{k=1}^{\infty} E_k\right) \cup \left(\bigcap_{k=2}^{\infty} E_k\right) \cup \left(\bigcap_{k=3}^{\infty} E_k\right) \cup \cdots \tag{2.223}$$

Suppose $x \notin \big\{x \in X : x \in E_k \text{ for all but finitely many } k\big\}$. That means, $\forall n \in \mathbb{N}$, $\exists j > n$ such that $x \notin \bigcap_{k=j}^{\infty} E_k$, which contradicts (2.223).

($\Leftarrow$) Let $x \in \liminf$ in (2.221). That means $\exists j_x$ such that

$$x \in \bigcap_{k=j_x}^{\infty} E_k \iff x \in \left(\bigcap_{k=1}^{\infty} E_k\right) \cup \left(\bigcap_{k=2}^{\infty} E_k\right) \cup \left(\bigcap_{k=3}^{\infty} E_k\right) \cup \cdots = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} E_k.$$

2.5.2 Tests for Convergence and Divergence

> With the exception of the geometric series, there does not exist in all of mathematics a single infinite series whose sum has been determined rigorously.
>
> (Niels Abel)

The following are some of the many conditions, or "tests", that can help determine if a series of nonnegative terms is convergent or divergent.

- **The *Comparison Test*** Let $S = \sum_{k=1}^{\infty} f_k$ and $T = \sum_{k=1}^{\infty} g_k$ with $0 \le f_k, g_k < \infty$ and T convergent. Then:

$$\textbf{If } \exists C > 0\textbf{: } f_k \le C g_k\textbf{, then } S \textbf{ converges.} \tag{2.224}$$

  This can be relaxed to requiring that that $f_k \le C g_k$ for all k sufficiently large, i.e., $\exists\, K \in \mathbb{N}$ such that it holds for all $k \ge K$.

  Proof: The proof when $f_k \le C g_k$ for all $k \in \mathbb{N}$ is simply to note that

$$\sum_{k=1}^{n} f_k \le \sum_{k=1}^{n} C g_k \le C \sum_{k=1}^{\infty} g_k < \infty$$

  is true for all n, so that the partial sum $\sum_{k=1}^{n} f_k$ is bounded. In a similar way, the comparison test can be used to show that a series diverges.

- **The *Ratio Test*** Let $S = \sum_{k=1}^{\infty} f_k$ with $0 \le f_k < \infty$.

$$\textbf{If } \exists c \in (0,1)\textbf{: } f_{k+1}/f_k \le c\textbf{, then } S \textbf{ converges.} \tag{2.225}$$

  This can be relaxed to stating that, if $\exists c \in (0,1)$ such that $f_{k+1}/f_k \le c$ for all k sufficiently large, then S converges.

  Proof: Let K be such that $f_{k+1} \le c f_k$ for all $k \ge K$. Then $f_{K+1} \le c f_K$, and $f_{K+2} \le c f_{K+1} \le c^2 f_K$, etc., and $f_{K+n} \le c^n f_K$. Then

$$\begin{aligned} \sum_{n=1}^{\infty} f_{K+n} &= f_{K+1} + f_{K+2} + \cdots \\ &\le c f_K + c^2 f_K + \cdots = \frac{c}{1-c} f_K, \end{aligned}$$

  which is finite for $c \in [0,1)$. More generally, allow f_k to be negative or positive, and let $c = \lim_{k\to\infty} |f_{k+1}/f_k|$. If $c < 1$, then a similar argument shows that $S = \sum_{k=1}^{\infty} f_k$ converges absolutely. If $c > 1$ or ∞, then $\exists K \in \mathbb{N}$ such that $\forall k \ge K$, $|f_k| > |f_K|$. This implies, from the contrapositive of (2.211), that $\lim_{k\to\infty} |f_k| \ne 0$, and S diverges.

- **The *Root Test*** Let $S = \sum_{k=1}^{\infty} f_k$ and $r = \lim_{k\to\infty} |f_k|^{1/k} \ge 0$.

$$\textbf{If } r < 1\textbf{, then } \sum_{k=1}^{\infty} |f_k| \textbf{ converges.} \tag{2.226}$$

Proof: If $r < 1$, then $\exists\, \epsilon > 0$ such that $r + \epsilon < 1$, and $\exists\, K \in \mathbb{N}$ such that $|f_k|^{1/k} < r + \epsilon$, or $|f_k| < (r+\epsilon)^k$, $\forall k \geq K$. It follows by the comparison test (2.224) with the geometric series $\sum_{k=1}^{\infty} (r+\epsilon)^k$ that $\sum_{k=1}^{\infty} |f_k|$ converges, i.e., S is absolutely convergent.

$$\textbf{If } r > 1 \textbf{ or } \infty\textbf{, then } S \textbf{ diverges}. \tag{2.227}$$

Proof: If $r > 1$ or ∞, then $\exists\, \epsilon > 0$ such that $r - \epsilon > 1$, and $\exists\, K \in \mathbb{N}$ such that $|f_k|^{1/k} > r - \epsilon$, or $|f_k| > (r-\epsilon)^k$, $\forall k \geq K$. Thus, $\lim_{k\to\infty} |f_k| > 1$, and S diverges.

$$\textbf{If } r = 1\textbf{, the test is inconclusive}. \tag{2.228}$$

Proof: Take the zeta function, with $f_k = k^{-p}$, and observe that, from the first limit result in Example 2.10,

$$\lim_{k\to\infty} f_k^{1/k} = \lim_{k\to\infty} \left(\frac{1}{k^p}\right)^{1/k} = \left(\frac{1}{\lim_{k\to\infty} k^{1/k}}\right)^p = 1$$

for any $p \in \mathbb{R}$. We know from Example 2.57 that $\zeta(p)$ converges for $p > 1$ and diverges otherwise, so that the root test is inconclusive.

- **The *Integral Test*** Let $f(x)$ be a nonnegative, decreasing function for all $x \geq 1$.

$$\textbf{If } \int_1^{\infty} f(x)\,dx < \infty\textbf{, then } S = \sum_{k=1}^{\infty} f(k) \textbf{ exists}. \tag{2.229}$$

Proof: This rests upon the fact that

$$f(k) \leq \int_{k-1}^{k} f(x)\,dx,$$

which is graphically obvious from Figure 9; the area of the rectangle from $n-1$ to n with height $f(n)$ is $1 \times f(n) = f(n)$, which is less than or equal to the area under $f(x)$ between $x = n-1$ and $x = n$. Thus, from the domain additivity property of the Riemann integral (2.138),

$$f(2) + f(3) + \cdots + f(k) \leq \int_1^k f(x)\,dx \leq \int_1^{\infty} f(x)\,dx < \infty,$$

and the partial sums are bounded. To show divergence, note from Figure 9 that $f(k) \geq \int_k^{k+1} f(x)\,dx$, so that

$$f(1) + f(2) + \cdots + f(k) \geq \int_1^{k+1} f(x)\,dx.$$

If the latter integral diverges as $k \to \infty$, then so does the partial sum.

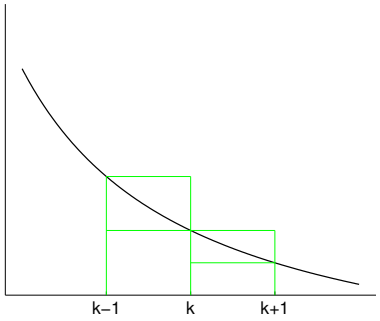


Figure 9: For continuous, positive, decreasing function f, $f(k) \leq \int_{k-1}^{k} f(x)dx$

- **The *Dirichlet Test*** Let $\{a_k\}$ and $\{b_k\}$ be sequences such that:
  - The partial sums of $\{a_k\}$ are bounded,
  - $\{b_k\}$ is positive and decreasing, i.e., $b_1 \geq b_2 \geq \cdots \geq 0$, and
  - $\lim_{k\to\infty} b_k = 0$.

  Then

$$\sum_{k=1}^{\infty} a_k b_k \textbf{ converges.} \tag{2.230}$$

  See, e.g., Stoll (2001, §7.2) for proof. As a special case, if f_k is a positive, decreasing sequence with $\lim_{k\to\infty} f_k = 0$, then $\sum_{k=1}^{\infty}(-1)^k f_k$ converges, which is often referred to as the **alternating series test**. Observe how the partial sums of $(-1)^k$ are bounded, but this sequence, and the sequence of partial sums, are not convergent.

  If the sequence $\{b_k\}$ is positive and decreasing, and $\lim_{k\to\infty} b_k = 0$, then the Dirichlet test can also be used to prove that $\sum_{k=1}^{\infty} b_k \sin(kt)$ converges for all $t \in \mathbb{R}$, and that $\sum_{k=1}^{\infty} b_k \cos(kt)$ converges for all $t \in \mathbb{R}$, except perhaps for $t = 2z\pi$, $z \in \mathbb{Z}$. See Stoll (2001, p. 296-7) for proof.

Example 2.60 *Let $f(x) = 1/x^p$ for $x \in \mathbb{R}_{\geq 1}$ and $p \in \mathbb{R}_{>0}$. As $f(x)$ is nonnegative and decreasing, the integral test (2.229) implies that $\zeta(p) = \sum_{x=1}^{\infty} x^{-p}$ exists if*

$$\int_1^{\infty} \frac{1}{x^p}\, dx = \lim_{x\to\infty}\left(\frac{x^{1-p}}{1-p}\right) - \frac{1}{1-p} < \infty,$$

which is true for $1 - p < 0$, i.e., $p > 1$, and does not exist otherwise. Thus, $\zeta(1)$ diverges, but $\zeta(p)$ converges for $p > 1$, as also stated in Example 2.57. ■

Example 2.61 *Let $S(p) = \sum_{k=1}^{\infty} (\ln k)/k^p$. For $p > 1$, use the "standard trick" (Lang, 1997, p. 212) and write $p = 1 + \epsilon + \delta$, $\delta > 0$. From (2.103) $\lim_{k\to\infty} (\ln k)/k^{\delta} = 0$, which implies that, for large enough k, $(\ln k)/k^{\delta} \leq 1$. Thus, for k large enough and $C = 1$,*

$$\frac{\ln k}{k^p} = \frac{\ln k}{k^{\delta} k^{1+\epsilon}} \leq C\frac{1}{k^{1+\epsilon}},$$

so that the comparison test (2.224) and the parameter range for convergence of the zeta function from Examples 2.57 and 2.60 imply that $S(p)$ converges for $p > 1$. A similar analysis shows that $S(p)$ diverges for $p < 1$. For $p = 1$, as $\ln k > 1$ for $k \geq 3$, the comparison test (2.224) with $\zeta(1)$ confirms that it also diverges. The integral test (2.229) also works in this case; see Stoll (2001, p. 286). ■

Example 2.62 *Continuing with the investigation of how inserting* $\ln k$ *affects convergence, consider now*

$$S(q) = \sum_{k=2}^{\infty} \frac{1}{(k \ln k)^q}.$$

First, take $q = 1$. *Let* $f : D \to \mathbb{R}$ *be given by* $f(x) = (x \ln x)^{-1}$, *with* $D = \{x \in \mathbb{R} : x \geq 2\}$. *It is straightforward to show that* $f'(x) = -(1 + \ln x)(x \ln x)^{-2} < 0$ *on* D, *so that, from the first derivative test,* f *is decreasing on* D, *and the integral test* (2.229) *is applicable. But from* (2.96), *the improper integral*

$$\lim_{t \to \infty} \int_2^t \frac{dx}{x \ln x} = \lim_{t \to \infty} [\ln(\ln t) - \ln(\ln 2)]$$

diverges, so that $S(1)$ *diverges.*[16] *For* $q > 1$, *let* $q = 1 + m$, *so that* (2.95) *with* $p = -m$ *implies that*

$$\int_2^t \frac{dx}{x(\ln x)^{1+m}} = -\frac{(\ln x)^{-m}}{m}\bigg|_2^t = \frac{(\ln 2)^{-m}}{m} - \frac{1}{m(\ln t)^m}, \quad m > 0,$$

so that

$$\int_2^{\infty} \frac{dx}{x(\ln x)^{1+m}} = \frac{(\ln 2)^{-m}}{m} < \infty,$$

and $S(q)$ *converges for* $q > 1$. ■

Example 2.63 *Let* $S = \sum_{k=2}^{\infty} (\ln k)^{-vk}$, *for constant* $v \in \mathbb{R}_{>0}$. *As*

$$\lim_{k \to \infty} \left|(\ln k)^{-vk}\right|^{1/k} = \lim_{k \to \infty} \left(\frac{1}{\ln k}\right)^v = 0,$$

the root test (2.226) *shows that* S *converges.* ■

Example 2.64 *Let*

$$\gamma_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} - \ln n,$$

which converges to Euler's constant, *denoted* γ, *with* $\gamma \approx 0.5772156649$. *As in Beardon (1997, p. 176), let*

$$a_n = \ln\left(\frac{n}{n-1}\right) - \frac{1}{n}.$$

That is, $a_2 = \ln 2 - \ln 1 - 1/2$, $a_3 = \ln 3 - \ln 2 - 1/3$, $a_4 = \ln 4 - \ln 3 - 1/4$, *etc., so the log terms in* $\sum_{i=2}^n a_i$ *are "telescoping". Thus,* $\sum_{i=2}^n a_i = \ln n - \ln 1 - 1/2 - 1/3 - \cdots - 1/n$, *i.e.,* $\sum_{i=2}^n a_i = 1 - \gamma_n$. *To see that* γ_n *is convergent, it suffices to show that* $\sum_{i=2}^{\infty} a_i$ *converges. Observe that (use substitution* $u = n - t$*)*

$$\int_0^1 \frac{t}{n(n-t)} dt = \int_0^1 \left(\frac{1}{n-t} - \frac{1}{n}\right) dt = a_n. \tag{2.231}$$

Next, let $f(t) = t/(n-t)$ *for* $n \geq 1$ *and* $t \in (0,1)$. *Clearly,* $f(0) = 0$, $f(t)$ *is increasing on* $(0,1)$, *and, as* $f''(t) = 2n(n-t)^{-3} > 0$, *it is convex. Thus, the area under its curve on*

[16] The divergence is clearly very slow. The largest number in Matlab is obtained by `t=realmax`, which is about 1.7977×10^{308}, and $\ln(\ln t) = 6.565$.

$(0,1)$ is bounded by that of the right triangle with vertices $(0,0)$, $(1,0)$ and $(1,f(1))$, which has area $f(1)/2 = \frac{1}{2(n-1)}$. Thus, from these results and the first integral in (2.231), $a_n \geq 0$. Further,

$$\int_0^1 \frac{t}{(n-t)} dt \leq \frac{1}{2(n-1)},$$

or

$$0 \leq a_n = \frac{1}{n}\int_0^1 \frac{t}{(n-t)} dt \leq \frac{1}{2n(n-1)} \leq \frac{1}{n^2}.$$

By the comparison test (2.224) with the zeta function result in Examples 2.57 and 2.60, $\sum_{i=2}^{\infty} a_i$ and, thus, γ_n, converge. See also the next example, and Example 2.84. ■

Example 2.65 *The following magnificent presentation is copied nearly verbatim from the (equally magnificent) Duren (Invitation to Classical Analysis, 2012, §2.3).*[17]

Euler's constant is

$$\gamma = \lim_{n\to\infty} \left\{ \sum_{k=1}^{n} \frac{1}{k} - \log n \right\}.$$

It is named for Leonhard Euler, who first discussed it in 1734. The number γ is an important constant that occurs frequently in mathematical formulas. The existence of the limit is not obvious. Our aim is to prove that the limit exists and to determine its approximate numerical value.

Consider the curve $y = 1/x$ for $1 \leq x \leq n$, where $n = 2, 3, \ldots$ The area under the curve is given by

$$A_n = \int_1^n \frac{1}{x} dx = \log n.$$

Now construct rectangular boxes of heights $1/k$ over the intervals $[k, k+1]$, as shown in the left panel of Figure 10.

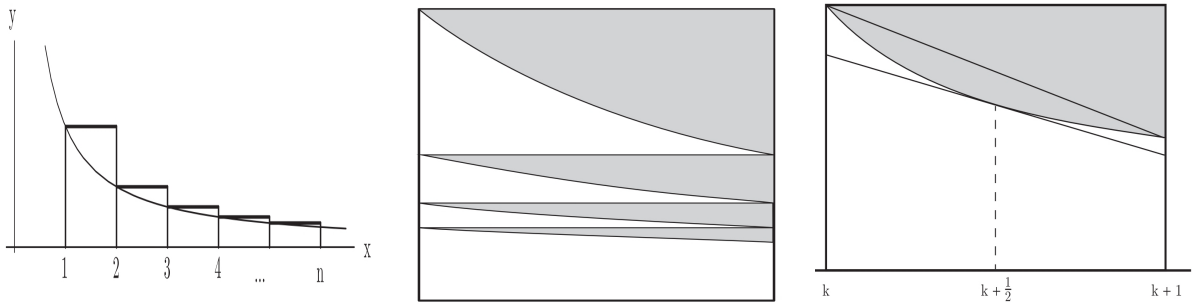


Figure 10: Left: The curve $y = 1/x$ and rectangular boxes. Middle: Geometric estimate of $S_{n-1} - A_n$. Right: Estimation of the area α_k.

Since

$$\frac{1}{k+1} \leq \frac{1}{x} \leq \frac{1}{k} \quad \text{for} \quad k \leq x \leq k+1,$$

it follows that

$$\frac{1}{k+1} = \int_k^{k+1} \frac{1}{k+1} dx \leq \int_k^{k+1} \frac{1}{x} dx \leq \int_k^{k+1} \frac{1}{k} dx = \frac{1}{k}$$

[17] On the other hand, Duren's approach, and useful graphics, shown below, are basically the same as those in Mattuck (Introduction to Analysis, 1999, p. 96).

for $k = 1, 2, \ldots$. Adding these inequalities over $k = 1, 2, \ldots, n-1$, we have

$$\sum_{k=1}^{n-1} \frac{1}{k+1} \leq \int_1^n \frac{1}{x} dx \leq \sum_{k=1}^{n-1} \frac{1}{k}.$$

With the notation

$$S_n = \sum_{k=1}^{n} \frac{1}{k}$$

this says that $S_n - 1 \leq A_n \leq S_{n-1}$. The two inequalities can be rearranged to give

$$0 \leq S_{n-1} - A_n \leq 1 - S_n + S_{n-1} = 1 - \frac{1}{n}.$$

This shows that the sequence $\{S_{n-1} - A_n\}$ is positive and is bounded above by 1.

Geometrically, the quantity $S_{n-1} - A_n$ is the sum of areas of those portions of the boxes that lie above the curve $y = 1/x$ from $x = 1$ to n. In order to estimate this total area, imagine that all of these boxes are slid to the left until they lie inside the first box, as shown in the middle panel of Figure 10, where the shaded regions have total area $S_{n-1} - A_n$. Since the regions are nonoverlapping and lie inside a square of area 1, this conceptual exercise gives a geometric interpretation of the inequality $S_{n-1} - A_n \leq 1$.

Next observe that

$$A_{n+1} - A_n = \int_n^{n+1} \frac{1}{x} dx \leq \frac{1}{n} = S_n - S_{n-1},$$

or $S_{n-1} - A_n \leq S_n - A_{n+1}$, which says that the sequence $\{S_{n-1} - A_n\}$ is nondecreasing. An appeal to the monotone boundedness theorem now shows that the sequence converges. Denoting its limit by γ, we have

$$\gamma = \lim_{n\to\infty} (S_{n-1} - A_n) = \lim_{n\to\infty} \left(S_n - \frac{1}{n} - A_n \right) = \lim_{n\to\infty} (S_n - A_n).$$

This establishes the existence of Euler's constant, as well as that $0 \leq \gamma \leq 1$.

In fact, it is clear from the middle panel of Figure 10 that γ is slightly larger than $1/2$. Our next task is to derive quantitative bounds on γ by estimating the area

$$\alpha_k = \frac{1}{k} - \int_k^{k+1} \frac{1}{x} dx$$

of the region in the k th box that lies above the curve $y = 1/x$. Since the curve is convex, that region contains a triangle of area

$$\frac{1}{2} \left(\frac{1}{k} - \frac{1}{k+1} \right),$$

and is contained in a trapezoid of area

$$\frac{1}{k} - \frac{1}{k + \frac{1}{2}},$$

constructed by drawing the tangent line to the curve at the point where $x = k + \frac{1}{2}$. (See the right panel of Figure 10, and note that the trapezoid above the tangent line is obtained by removing the lower trapezoid from the entire rectangle.)

Thus a comparison of areas shows that

$$\frac{1}{2}\left(\frac{1}{k}-\frac{1}{k+1}\right)\leq\alpha_k\leq\frac{1}{k}-\frac{1}{k+\frac{1}{2}}=2\left(\frac{1}{2k}-\frac{1}{2k+1}\right).$$

Summing these inequalities, we find

$$\frac{1}{2}\sum_{k=1}^{n-1}\left(\frac{1}{k}-\frac{1}{k+1}\right)\leq\sum_{k=1}^{n-1}\alpha_k\leq 2\sum_{k=1}^{n-1}\left(\frac{1}{2k}-\frac{1}{2k+1}\right),$$

or

$$\frac{1}{2}\left(1-\frac{1}{n}\right)\leq S_{n-1}-A_n\leq 2\left(\frac{1}{2}-\frac{1}{3}+\frac{1}{4}-\frac{1}{5}+\cdots-\frac{1}{2n-1}\right).$$

Letting $n\to\infty$, we infer that

$$0.5\leq\gamma\leq 2\left(\frac{1}{2}-\frac{1}{3}+\frac{1}{4}-\frac{1}{5}+\ldots\right)=2(1-\log 2)=0.6137\ldots,$$

since, as shown below in (2.278),

$$1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\cdots=\log 2.$$

The actual value of Euler's constant is

$$\gamma=0.577215664901532\ldots.$$

It has been computed to thousands of decimal places and no periodicities have been detected, so it is strongly suspected to be an irrational number. In fact, it is generally conjectured to be a transcendental (or nonalgebraic) number, like the constants π and e. However, no one has ever been able to prove that γ is irrational! ■

Example 2.66 *To see that $\cos(z)=\sum_{k=0}^{\infty}(-1)^k z^{2k}/(2k)!$ converges, use the ratio test (2.225) to see that*

$$c=\lim_{k\to\infty}\left|\frac{(-1)^{k+1}\frac{z^{2(k+1)}}{(2(k+1))!}}{(-1)^k\frac{z^{2k}}{(2k)!}}\right|=\lim_{k\to\infty}\left|\frac{z^2}{(2k+1)(2k+2)}\right|=0,$$

for all $z\in\mathbb{R}$. ■

Example 2.67 *Consider the sum $S=\lim_{r\to\infty}\sum_{i=1}^{r-1}\frac{i}{(r-i)^2}$ and let $j=r-i$ so that*

$$S=\lim_{r\to\infty}\sum_{j=1}^{r}\frac{r-j}{j^2}=\lim_{r\to\infty}\left(r\sum_{j=1}^{r}\frac{1}{j^2}-\sum_{j=1}^{r}\frac{1}{j}\right).$$

Then

$$\frac{S}{r}=\lim_{r\to\infty}\sum_{j=1}^{r}\frac{1}{j^2}-\lim_{r\to\infty}\frac{1}{r}\sum_{j=1}^{r}\frac{1}{j}. \tag{2.232}$$

From (2.217), the first sum in (2.232) converges to $\pi^2/6$. Using the comparison test (2.224), the second sum is bounded because, for $r \geq 2$,

$$\sum_{j=1}^{r} \frac{1}{rj} < \sum_{j=1}^{r} \frac{1}{r} = 1.$$

To see that it converges to zero, use the fact that $(rj)^{-1}$ is a positive, decreasing function in j, so that the conditions of the integral test (2.229) hold. Then,

$$\sum_{j=2}^{r} \frac{1}{rj} \leq \int_1^r \frac{1}{rx} dx \qquad \text{and} \qquad \lim_{r\to\infty} \int_1^r \frac{1}{rx} dx = \lim_{r\to\infty} \frac{\ln r}{r} = \lim_{r\to\infty} \frac{r^{-1}}{1} = 0$$

from l'Hôpital's rule. Thus, S/r in (2.232) converges to $\pi^2/6$ and, as $r \to \infty$,

$$S \to \frac{r\pi^2}{6},$$

in the sense that the ratio of S to $r\pi^2/6$ converges to unity as r grows. This result is used in Paolella, Fundamental Probability, Example 6.8, for the calculation of the asymptotic variance of a particular random variable. ■

Of great importance in analysis is knowing the conditions under which one can exchange limiting operations. For example, in §5.3, we consider the exchange of derivative and integral. Another case is the exchange of a limit and an infinite sum. We give here *Tannery's theorem (for series)*, after Jules Tannery (1848–1910), which is often of great use, as we will show with examples. It turns out to be a special case of the famous Lebesgue dominated convergence theorem (DCT), the latter requiring a study of measure theory and the Lebesgue integral to understand. In fact, Tannery's theorem for series is sometimes referred to as the discrete DCT. Fortunately, Tannery's theorem can be proven without invoking this machinery, and is in fact rather straightforward. The following presentation is taken from Loya (Amazing and Aesthetic Aspects of Analysis, 2018, §3.7.1).

Theorem (Tannery's theorem for series): For each natural number n, let $\sum_{k=1}^{m_n} a_k(n)$ be a finite sum such that $m_n \to \infty$ as $n \to \infty$. If for each k, $\lim_{n\to\infty} a_k(n)$ exists, and there is a convergent series $\sum_{k=1}^{\infty} M_k$ of nonnegative real numbers such that $|a_k(n)| \leq M_k$ for all $n \in \mathbb{N}$ and $1 \leq k \leq m_n$, then

$$\lim_{n\to\infty} \sum_{k=1}^{m_n} a_k(n) = \sum_{k=1}^{\infty} \lim_{n\to\infty} a_k(n). \tag{2.233}$$

In particular, both sides are well defined (the limits and sums converge) and are equal.

> Proof: First of all, we remark that the series on the right converges. Indeed, if we put $a_k := \lim_{n\to\infty} a_k(n)$ (the limit exists by assumption), then taking $n \to \infty$ in the inequality $|a_k(n)| \leq M_k$, we have $|a_k| \leq M_k$ as well. Therefore, by the comparison test, $\sum_{k=1}^{\infty} |a_k|$ converges, and hence $\sum_{k=1}^{\infty} a_k$ converges as well.
>
> Now to prove our theorem, let $\varepsilon > 0$ be given. It follows from Cauchy's criterion for series that there is an ℓ such that
>
> $$M_{\ell+1} + M_{\ell+2} + \cdots < \frac{\varepsilon}{3}.$$

Since $m_n \to \infty$ as $n \to \infty$, we can choose N_1 such that for all $n > N_1$, we have $m_n > \ell$. Then using that $|a_k(n)| \leq M_k$ and $|a_k| \leq M_k$, observe that for every $n > N_1$,

$$\begin{aligned}\left|\sum_{k=1}^{m_n} a_k(n) - \sum_{k=1}^{\infty} a_k\right| &= \left|\sum_{k=1}^{\ell} (a_k(n) - a_k) + \sum_{k=\ell+1}^{m_n} a_k(n) - \sum_{k=\ell+1}^{\infty} a_k\right| \\ &\leq \sum_{k=1}^{\ell} |a_k(n) - a_k| + \sum_{k=\ell+1}^{m_n} M_k + \sum_{k=\ell+1}^{\infty} M_k \\ &< \sum_{k=1}^{\ell} |a_k(n) - a_k| + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \sum_{k=1}^{\ell} |a_k(n) - a_k| + \frac{2\varepsilon}{3}.\end{aligned}$$

Since for each $k, \lim_{n\to\infty} a_k(n) = a_k$, there is an N_2 such that for each $k = 1, 2, \ldots, \ell$ and for $n > N_2$, we have $|a_k(n) - a_k| < \varepsilon/(3\ell)$. Thus, if $n > \max\{N_1, N_2\}$, then

$$\left|\sum_{k=1}^{m_n} a_k(n) - \sum_{k=1}^{\infty} a_k\right| < \sum_{k=1}^{\ell} \frac{\varepsilon}{3\ell} + \frac{2\varepsilon}{3} = \frac{\varepsilon}{3} + \frac{2\varepsilon}{3} = \varepsilon.$$

This completes the proof.

Our first example is a "non-example", showing that the necessity of having a convergent dominating series.

Example 2.68 *(Loya, p. 218) For each $k, n \in \mathbb{N}$, let $a_k(n) := 1/n$ and let $m_n = n$. Then*

$$\lim_{n\to\infty} a_k(n) = \lim_{n\to\infty} \frac{1}{n} = 0 \Longrightarrow \sum_{k=1}^{\infty} \lim_{n\to\infty} a_k(n) = \sum_{k=1}^{\infty} 0 = 0.$$

On the other hand,

$$\sum_{k=1}^{m_n} a_k(n) = \sum_{k=1}^{n} \frac{1}{n} = \frac{1}{n} \cdot \sum_{k=1}^{n} 1 = 1 \Longrightarrow \lim_{n\to\infty} \sum_{k=1}^{m_n} a_k(n) = \lim_{n\to\infty} 1 = 1$$

Thus, for this example,

$$\lim_{n\to\infty} \sum_{k=1}^{m_n} a_k(n) \neq \sum_{k=1}^{\infty} \lim_{n\to\infty} a_k(n)$$

It turns out there is no constant M_k such that $|a_k(n)| \leq M_k$ where the series $\sum_{k=1}^{\infty} M_k$ converges. Indeed, the inequality $|a_k(n)| = 1/n \leq M_k$ for $1 \leq k \leq n$ implies (set $k = n$) that $1/k \leq M_k$ for all k. Since $\sum_{k=1}^{\infty} 1/k$ diverges, the series $\sum_{k=1}^{\infty} M_k$ must also diverge. ■

Example 2.69 *(Wikipedia, Tannery's Theorem) We wish to prove that the limit of the binomial theorem (1.21), and the infinite series characterizations of the exponential e^x (2.241), are equivalent. Note that*

$$\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n = \lim_{n\to\infty} \sum_{k=0}^{n} \binom{n}{k} \frac{x^k}{n^k}.$$

Define $a_k(n) = \binom{n}{k}\frac{x^k}{n^k}$. We have that $|a_k(n)| \leq \frac{|x|^k}{k!}$ and that $\sum_{k=0}^{\infty} \frac{|x|^k}{k!} = e^{|x|} < \infty$, so Tannery's theorem can be applied and

$$\lim_{n\to\infty} \sum_{k=0}^{\infty} \binom{n}{k} \frac{x^k}{n^k} = \sum_{k=0}^{\infty} \lim_{n\to\infty} \binom{n}{k} \frac{x^k}{n^k} = \sum_{k=0}^{\infty} \frac{x^k}{k!} = e^x.$$ ■

Another example of the use of Tannery's theorem is given in Example 6.3 in the context of the digamma function.

Definition: The *exponential growth rate* of the series $\sum_{k=1}^{\infty} f_k$ is given by $L = \limsup |f_k|^{1/k}$.

Theorem: In the root test (2.226), the assumption that $\lim_{k\to\infty} |f_k|^{1/k}$ exists can be relaxed by working with the exponential growth rate of the series $\sum_{k=1}^{\infty} f_k$, which always exists (in $\mathbb{X}$).

> Proof: This is similar to the proof of the root test. If $L < 1$, then $\exists \epsilon > 0$ such that $L + \epsilon < 1$, and, from (2.203), $\exists K \in \mathbb{N}$ such that, $\forall k \geq K$, $|f_k|^{1/k} < L + \epsilon$, or $|f_k| < (L + \epsilon)^k$. The comparison test is then used as before. A similar argument using (2.204) shows that the series diverges if $L > 1$.

Example 2.70 *(Stoll, 2001, p. 289) Let* $a_n = 2^{-k}$ *if* $n = 2k$ *and* $a_n = 3^{-k}$ *if* $n = 2k + 1$*, so that*

$$S = \sum_{n=1}^{\infty} a_n = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{2^2} + \frac{1}{3^2} + \cdots,$$

and $a_n^{1/n} = 2^{-1/2}$ *for* $n = 2k$ *and* $a_n^{1/n} = 3^{-k/(2k+1)}$ *for* $n = 2k + 1$*. As* $\lim_{k\to\infty} 3^{-k/(2k+1)} = 3^{-1/2}$ *and* $\max\left(2^{-1/2}, 3^{-1/2}\right) = 2^{-1/2}$*, we have* $\limsup a_n^{1/n} = 2^{-1/2} < 1$ *and, thus, by the root test,* S *converges.* ■

Remarks:

(a) It turns out that the root test is "more powerful" than the ratio test, in the sense that, if the ratio test proves convergence, or divergence, then so does the root test, but the converse is not true. (Our comfortable statistics language is not actually used; one says that the root test has a *strictly wider scope* than the ratio test.) The reason for this is that, for the positive sequence $\{a_n\}$,

$$\liminf \frac{a_{n+1}}{a_n} \leq \liminf a_n^{1/n} \leq \limsup a_n^{1/n} \leq \limsup \frac{a_{n+1}}{a_n}. \tag{2.234}$$

The proof of this is straightforward; see, e.g., Stoll (2001, p. 291). It is, however, not hard to construct series for which the ratio test is much easier to apply than the root test, so that both tests are indeed valuable.

(b) For the series $S = \sum_{k=1}^{\infty} f_k$, if $L = \limsup |f_k|^{1/k} < 1$, then S is absolutely convergent, and hence also convergent. If $L > 1$, then $\sum_{k=1}^{\infty} |f_k|$ diverges, but could it be the case that S converges?

To answer this, recall from (2.204) that, if $L > 1$, then there are infinitely many k such that $|f_k| > 1$. Thus, whatever the sign of the terms, $\{f_k\}$ is not converging to zero, which implies that S diverges. ■

The tools for convergence of infinite series can be extended to convergence of *infinite products* by use of basic properties of the logarithm and continuity of the exponential function (see §2.2.4).

Theorem: Let $a_k \geq 0$ such that $\lim_{k\to\infty} a_k = 0$. Then:

$$\textbf{If } S = \sum_{k=1}^{\infty} a_k \textbf{ converges, then so does } P = \prod_{k=1}^{\infty} (1 + a_k). \tag{2.235}$$

Proof: Let $p_n = \prod_{k=1}^{n} (1 + a_k)$. As $\ln(1 + a_k) \leq a_k$,

$$\ln p_n = \sum_{k=1}^{n} \ln(1 + a_k) \leq \sum_{k=1}^{n} a_k \leq S,$$

so that, from the comparison test, $\ln P = \sum_{k=1}^{\infty} \ln(1 + a_k)$ converges. Taking exponents gives the result.

2.5.3 Wallis' Product and Stirling's Approximation

Wallis' formula, or *Wallis' product*, after John Wallis (1616–1703), is not only of interest in itself, but useful when deriving Stirling's approximation, both of which we detail in this subsection. As a bit of trivia, the sideways eight symbol, ∞, was introduced in 1655 by Wallis.

From (2.104), $\lim_{n\to\infty} (1 - t/n)^n = e^{-t}$. Following Keeping (1995, p. 392), for $x > 0$ and using $u = t/n$ and (1.51), $\Gamma(x)$ is given by

$$\lim_{n\to\infty} \int_0^n \left(1 - \frac{t}{n}\right)^n t^{x-1} dt = \lim_{n\to\infty} n^x \int_0^1 u^{x-1} (1 - u)^n \, du = \lim_{n\to\infty} n^x \frac{\Gamma(x)\,\Gamma(n+1)}{\Gamma(x + n + 1)}.$$

Dividing by $\Gamma(x)$ gives

$$1 = \lim_{n\to\infty} \frac{n^x \Gamma(n+1)}{\Gamma(x + n + 1)}.$$

But with $x = 1/2$ and using (1.39) and (1.44),

$$\begin{aligned}
\frac{n^x \Gamma(n+1)}{\Gamma(x + n + 1)} &= \frac{n^{1/2} n!}{\left(n + \frac{1}{2}\right)\left(n - \frac{1}{2}\right)\left(n - \frac{3}{2}\right)\cdots \frac{1}{2}\Gamma\left(\frac{1}{2}\right)} = \frac{n^{1/2} n!}{\left(\frac{2n+1}{2}\right)\left(\frac{2n-1}{2}\right)\left(\frac{2n-3}{2}\right)\cdots \frac{1}{2}\sqrt{\pi}} \\
&= \frac{n^{1/2} n!}{\frac{\left(\frac{2n+1}{2}\right)\left(\frac{2n}{2}\right)\left(\frac{2n-1}{2}\right)\left(\frac{2n-2}{2}\right)\left(\frac{2n-3}{2}\right)\cdots \frac{1}{2}\sqrt{\pi}}{\left(\frac{2n}{2}\right)\left(\frac{2n-2}{2}\right)\cdots 1}} = \frac{n^{1/2} n!}{\frac{(2n+1)!\sqrt{\pi}}{2^{2n+1} n!}} = \frac{2^{2n+1} n^{1/2} (n!)^2}{(2n+1)!\sqrt{\pi}},
\end{aligned}$$

or

$$\begin{aligned}
\sqrt{\pi} &= \lim_{n\to\infty} \frac{n^{1/2} 2^{2n+1} (n!)^2}{(2n+1)!} = \lim_{n\to\infty} \frac{1}{n^{1/2}} \frac{2n}{(2n+1)} \frac{2^{2n} (n!)^2}{(2n)!} \\
&= \lim_{n\to\infty} \frac{1}{n^{1/2}} \frac{2^{2n} (n!)^2}{(2n)!} = \lim_{n\to\infty} \frac{1}{n^{1/2}} \frac{(2n)^2 (2n-2)^2 (2n-4)^2 \cdots}{(2n)(2n-1)(2n-2)(2n-3)\cdots} \\
&= \lim_{n\to\infty} \frac{1}{n^{1/2}} \frac{(2n)(2n-2)(2n-4)\cdots 2}{(2n-1)(2n-3)\cdots 1},
\end{aligned}$$

which is *Wallis' product*,

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} \frac{2 \cdot 4 \cdot 6 \cdots 2n}{1 \cdot 3 \cdot 5 \cdots (2n-1)} = \sqrt{\pi}. \tag{2.236}$$

Theorem: Wallis' product can also be expressed as

$$\lim_{n\to\infty} \frac{2}{1}\frac{2}{3}\frac{4}{3}\frac{4}{5}\frac{6}{5}\frac{6}{7} \cdots \frac{2n}{2n-1}\frac{2n}{2n+1} = \frac{\pi}{2}. \tag{2.237}$$

This is proven below. Before doing so, we first take care of two other things. To see that the product in (2.237) converges, use the infinite product result (2.235) with

$$\frac{2n}{2n-1}\frac{2n}{2n+1} = \frac{4n^2}{4n^2-1} = 1 + \frac{1}{4n^2-1} =: 1 + a_k.$$

To show that $\sum_{k=1}^{\infty} a_k$ converges, use the comparison test (2.224) with $\zeta(2)$.

Next consider how to obtain (2.236) from (2.237). As in Loya, express the latter as

$$\frac{\pi}{2} = \lim_{n\to\infty}\left\{\left(\frac{2}{1}\right)^2 \cdot \left(\frac{4}{3}\right)^2 \cdots \left(\frac{2n}{2n-1}\right)^2 \cdot \frac{1}{2n+1}\right\}.$$

Then taking square roots,

$$\sqrt{\pi} = \lim_{n\to\infty}\sqrt{\frac{2}{2n+1}}\prod_{k=1}^{n}\frac{2k}{2k-1} = \lim_{n\to\infty}\frac{1}{\sqrt{n}}\frac{1}{\sqrt{1+1/2n}}\prod_{k=1}^{n}\frac{2k}{2k-1}.$$

Using that $1/\sqrt{1+1/2n} \to 1$ as $n \to \infty$, we obtain (2.236).

There are several ways of proving (2.237); see, e.g., Keeping (1995, p. 392), Andrews, Askey and Roy (1999, p. 46), and the (charming and excellent) Loya (Amazing and Aesthetic Aspects of Analysis, 2018, §5.1.3). We present the approach as in Hijab (Introduction to Calculus and Classical Analysis, 4th ed, 2016, p 204-5).

Proof: Begin with integrating by parts to obtain

$$\int \sin^n x dx = -\frac{1}{n}\sin^{n-1} x \cos x + \frac{n-1}{n}\int \sin^{n-2} x dx, \quad n \geq 2.$$

Evaluating at 0 and $\pi/2$ yields

$$\int_0^{\pi/2} \sin^n x dx = \frac{n-1}{n}\int_0^{\pi/2} \sin^{n-2} x dx, \quad n \geq 2.$$

Since $\int_0^{\pi/2} \sin^0 x dx = \pi/2$ and $\int_0^{\pi/2} \sin^1 x dx = 1$, by the last equation and induction,

$$I_{2n} = \int_0^{\pi/2} \sin^{2n} x dx = \frac{(2n-1)\cdot(2n-3)\cdot\cdots\cdot 1}{2n\cdot(2n-2)\cdot\cdots\cdot 2}\cdot\frac{\pi}{2},$$

and

$$I_{2n+1} = \int_0^{\pi/2} \sin^{2n+1} x dx = \frac{2n\cdot(2n-2)\cdot\cdots\cdot 2}{(2n+1)\cdot(2n-1)\cdot\cdots\cdot 3}\cdot 1,$$

for $n \geq 1$. Since $0 < \sin x < 1$ on $(0, \pi/2)$, the integrals I_n are decreasing in n. But, by the formula for I_n with n odd,

$$1 \leq \frac{I_{2n-1}}{I_{2n+1}} \leq 1 + \frac{1}{2n}, \quad n \geq 1.$$

Thus

$$1 \leq \frac{I_{2n}}{I_{2n+1}} \leq \frac{I_{2n-1}}{I_{2n+1}} \leq 1 + \frac{1}{2n}, \quad n \geq 1,$$

or $I_{2n}/I_{2n+1} \to 1$, as $n \to \infty$. Since

$$\frac{I_{2n}}{I_{2n+1}} = \frac{(2n+1)\cdot(2n-1)\cdot(2n-1)\cdot\cdots\cdot 3\cdot 3\cdot 1}{2n\cdot 2n\cdot(2n-2)\cdot\cdots\cdot 4\cdot 2\cdot 2}\cdot\frac{\pi}{2},$$

we obtain (2.237).

As mentioned, the Wallis product is important for deriving *Stirling's approximation* to $n!$,

$$\Gamma(n) \approx \sqrt{2\pi} n^{n-1/2} \exp(-n). \tag{2.238}$$

The derivation of this famous result is often given in books on real analysis; e.g., Lang (1997, p. 120), Andrews, Askey and Roy (1999, §1.4), Kuttler (2021, Calculus of One and Many Variables, §10.1), and Duren (2012, §2.6). Our presentation is taken from Duren.

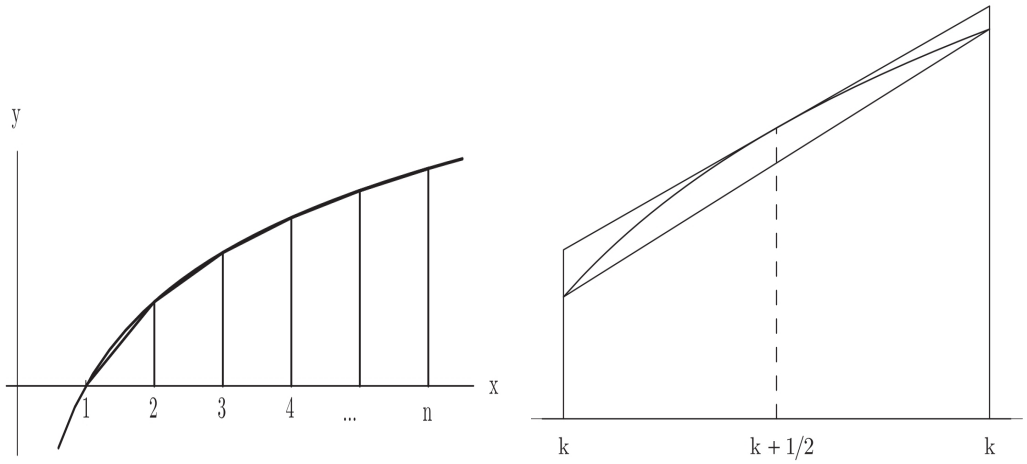


Figure 11: Left: The logarithmic curve and inscribed trapezoids. Right: Estimation of the area α_k.

The asymptotic formula

$$n! \sim n^n e^{-n} \sqrt{2\pi n}, \quad n \to \infty,$$

is known as Stirling's formula.[18] It is of basic importance for instance in probability theory and combinatorics, because it gives precise information about the growth of the factorial function. The symbol "$\sim$" means that

$$\lim_{n\to\infty} \frac{n!}{n^n e^{-n}\sqrt{2\pi n}} = 1.$$

We propose to prove Stirling's formula by showing that

$$1 < \frac{n!}{n^n e^{-n}\sqrt{2\pi n}} < 1 + \frac{1}{4n}, \quad n = 1, 2, \ldots.$$

The error estimates are important in quantitative applications of the formula.

With the observation that

$$\log(n!) = \log 1 + \log 2 + \cdots + \log n$$

it is natural to base a proof on a careful study of the area under the logarithmic curve $y = \log x$ from $x = 1$ to n. (See the left panel of Figure 11.) An integration by parts calculates this area as

$$A_n = \int_1^n \log x dx = [x \log x - x]_1^n = n \log n - n + 1.$$

On the other hand, the area can be estimated geometrically. Since the logarithm is a concave function, the curve $y = \log x$ lies above each of its chords connecting successive points $(k, \log k)$, for $k = 1, 2, \ldots, n$. Thus A_n is larger than the sum of areas of the trapezoids under those line segments. The total area of the trapezoids is

$$\begin{aligned} T_n &= \frac{1}{2}\log 2 + \frac{1}{2}(\log 2 + \log 3) + \cdots + \frac{1}{2}(\log(n-1) + \log n) \\ &= \log 2 + \log 3 + \cdots + \log(n-1) + \frac{1}{2}\log n \\ &= \log(n!) - \frac{1}{2}\log n. \end{aligned}$$

Now let α_k denote the area of the small region bounded by the curve $y = \log x$ and the line segment joining the two points $(k, \log k)$ and $(k+1, \log(k+1))$, for $k = 1, 2, \ldots, n-1$. Then the total area under the curve is

$$A_n = T_n + E_n, \qquad \text{where } E_n = \alpha_1 + \alpha_2 + \cdots + \alpha_{n-1}.$$

Inserting the expressions for A_n and T_n, we can write this relation in the form

$$\log(n!) = \left(n + \frac{1}{2}\right)\log n - n + 1 - E_n,$$

or

$$n! = C_n n^{n+\frac{1}{2}} e^{-n}, \qquad \text{where } C_n = e^{1-E_n}.$$

The sequence $\{E_n\}$ is increasing, since each term α_k is positive. We now show that the sequence $\{E_n\}$ has an upper bound and is therefore convergent. In order to estimate α_k, we construct the tangent line to the curve $y = \log x$ at the point where $x = k + \frac{1}{2}$ (see the right panel of Figure 11) and compare areas:

$$\begin{aligned} \alpha_k &< \log\left(k + \frac{1}{2}\right) - \frac{1}{2}(\log k + \log(k+1)) \\ &= \frac{1}{2}\log\left(\frac{k+\frac{1}{2}}{k}\right) - \frac{1}{2}\log\left(\frac{k+1}{k+\frac{1}{2}}\right) \\ &= \frac{1}{2}\log\left(1 + \frac{1}{2k}\right) - \frac{1}{2}\log\left(1 + \frac{1}{2k+1}\right) \\ &< \frac{1}{2}\log\left(1 + \frac{1}{2k}\right) - \frac{1}{2}\log\left(1 + \frac{1}{2(k+1)}\right). \end{aligned}$$

Adding these inequalities, we find that

$$\begin{aligned} E_n = \sum_{k=1}^{n-1} \alpha_k &< \sum_{k=1}^{n-1} \left\{ \frac{1}{2}\log\left(1 + \frac{1}{2k}\right) - \frac{1}{2}\log\left(1 + \frac{1}{2(k+1)}\right) \right\} \\ &= \frac{1}{2}\log\frac{3}{2} - \frac{1}{2}\log\left(1 + \frac{1}{2n}\right) < \frac{1}{2}\log\frac{3}{2}, \end{aligned}$$

since the dominant series telescopes. Thus E_n increases to a finite limit $E = \sum_{k=1}^{\infty} \alpha_k$, and so $C_n = e^{1-E_n}$ decreases to a limit $C = e^{1-E} > 0$. In particular, $C_n > C$, and so $1 < C_n/C = e^{E-E_n}$. But

$$\begin{aligned} E - E_n = \sum_{k=n}^{\infty} \alpha_k &< \sum_{k=n}^{\infty} \left\{ \frac{1}{2}\log\left(1 + \frac{1}{2k}\right) - \frac{1}{2}\log\left(1 + \frac{1}{2(k+1)}\right) \right\} \\ &= \frac{1}{2}\log\left(1 + \frac{1}{2n}\right), \end{aligned}$$

again because the dominant series telescopes. Therefore,

$$1 < C_n/C = e^{E-E_n} < \sqrt{1+\frac{1}{2n}} < 1+\frac{1}{4n}.$$

In summary, we have shown that

$$0 < C < C_n = \frac{n!}{n^{n+\frac{1}{2}}e^{-n}} < C\left(1+\frac{1}{4n}\right).$$

In order to finish the proof of Stirling's formula with error estimates, it now remains only to show that $C = \sqrt{2\pi}$. This is where we invoke the Wallis product formula. It gives

$$\begin{aligned}\sqrt{\pi} &= \lim_{n\to\infty} \frac{2^{2n}(n!)^2}{(2n)!\sqrt{n}} = \lim_{n\to\infty} \frac{2^{2n}\left(C_n n^{n+\frac{1}{2}}e^{-n}\right)^2}{\left(C_{2n}(2n)^{2n+\frac{1}{2}}e^{-2n}\right)\sqrt{n}} \\ &= \frac{1}{\sqrt{2}} \lim_{n\to\infty} \frac{C_n^2}{C_{2n}} = \frac{C^2}{\sqrt{2}C} = \frac{C}{\sqrt{2}}.\end{aligned}$$

Thus $C = \sqrt{2\pi}$ and the proof is complete.

Example 2.71 *We now provide a vastly faster and easier, albeit heuristic, derivation of Stirling's approximation using probability theory. Let $S_n \sim \mathrm{Gam}(n,1)$ for $n \in \mathbb{N}$, so that, for large n, $S_n \overset{app}{\sim} \mathrm{N}(n,n)$. The definition of convergence in distribution, and the continuity of the c.d.f. of S_n and that of its limiting distribution, informally suggest the limiting behavior of the p.d.f. of S_n, i.e.,*

$$f_{S_n}(s) = \frac{1}{\Gamma(n)} s^{n-1}\exp(-s) \approx \frac{1}{\sqrt{2\pi n}} \exp\left(-\frac{(s-n)^2}{2n^2}\right).$$

Choosing $s = n$ leads to $\Gamma(n+1) = n! \approx \sqrt{2\pi}(n+1)^{n+1/2}\exp(-n-1)$. From (2.104), $\lim_{n\to\infty}(1+\lambda/n)^n = e^\lambda$, so

$$(n+1)^{n+1/2} = n^{n+1/2}\left(1+\frac{1}{n}\right)^{n+1/2} \approx n^{n+1/2}e,$$

and substituting this into the previous expression for $n!$ yields Stirling's approximation $n! \approx \sqrt{2\pi}n^{n+1/2}e^{-n}$.

As an aside, Stirling's approximation also drops out of an application of a saddlepoint approximation; see Paolella, Intermediate Probability. ■

2.5.4 Cauchy Product

> Cauchy is mad and there is nothing that can be done about him, although, right now, he is the only one who knows how mathematics should be done.
>
> (Niels Abel)

Consider the product of the two series $\sum_{k=0}^{\infty} a_k$ and $\sum_{k=0}^{\infty} b_k$. Multiplying their values

out in tabular form

	b_0	b_1	b_2	$\cdots$
a_0	a_0b_0	a_0b_1	a_0b_2	$\cdots$
a_1	a_1b_0	a_1b_1	a_1b_2	
a_2	a_2b_0	a_2b_1	a_2b_2	
$\vdots$	$\vdots$			$\ddots$

and summing the off–diagonals suggests that the product is given by the

$$a_0b_0 + (a_0b_1 + a_1b_0) + (a_0b_2 + a_1b_1 + a_2b_0) + \cdots,$$

which is referred to as the *Cauchy product*. It is a standard result in real analysis that, if $\sum_{k=0}^{\infty} a_k$ and $\sum_{k=0}^{\infty} b_k$ are absolutely convergent series with sums A and B, respectively, then their Cauchy product

$$\sum_{k=0}^{\infty} c_k, \qquad c_k = a_0b_k + a_1b_{k-1} + \cdots + a_kb_0, \tag{2.239}$$

is absolutely convergent with sum AB (see, e.g., Trench, 2003, p. 227).

Example 2.72 *Let* $S = \sum_{k=0}^{\infty} a^k = (1-a)^{-1}$ *for* $a \in [0,1)$. *As this is absolutely convergent, (2.239) with* $a_k = b_k = a^k$ *implies that* $c_k = a^0a^k + a^1a^{k-1} + \cdots + a^ka^0 = (k+1)\,a^k$ *and* $(1-a)^{-2} = S^2 = \sum_{k=0}^{\infty} c_k = 1 + 2a + 3a^2 + \cdots$. ■

The Cauchy product result can be generalized. Let $x_{nm} := x(n,m)$ be a function of $n, m \in \mathbb{N}$ with $x_{nm} \in \mathbb{R}_{\geq 0}$. Then

$$\lim_{N\to\infty} \sum_{n=0}^{N}\sum_{m=0}^{N} x_{nm} = \lim_{N\to\infty} \sum_{m=0}^{N}\sum_{n=0}^{N} x_{nm} = \sum_{s=0}^{\infty} \sum_{\substack{m\geq 0, n\geq 0 \\ m+n=s}} x_{nm}, \tag{2.240}$$

if the unordered sum converges or, equivalently, if the terms are absolutely summable (see e.g., Beardon, 1997, §5.5; Browder, 1996, §2.5). As before, the values to be summed can be shown in a table as

		n			
		0	1	2	$\cdots$
	0	x_{00}	x_{01}	x_{02}	$\cdots$
m	1	x_{10}	x_{11}	x_{12}	
	2	x_{20}	x_{21}	x_{22}	
	$\vdots$	$\vdots$			$\ddots$

and, if the double sum is absolutely convergent, then the elements can be summed in any order, i.e., by columns, by rows, or by summing the off-diagonals.

Example 2.73 *It is instructive to show some of the properties of the exponential starting from its power series expression*

$$f(x) = \exp(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots. \tag{2.241}$$

Most students are familiar with this from their basic calculus class, and we will justify this rigorously below in §2.5.6. In particular, it is easily determined from (2.282) with $c = 0$. Clearly, $f(0) = 1$, and observe that, for all $x \in \mathbb{R}$,

$$\lim_{k\to\infty} \left| \frac{x^{k+1}/(k+1)!}{x^k/k!} \right| = \lim_{k\to\infty} \left| \frac{x}{k+1} \right| = 0,$$

so that the ratio test shows absolute convergence of the series. Differentiating (2.241) termwise[19] *shows that $f(x) = f'(x)$. Thus, from (2.78) and (2.79), with $s_n(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$, $s_n(x) \to \exp(x)$ for all $x \in \mathbb{R}$.*

Next, to show $f(x+y) = f(x) f(y)$ from (2.241), use the binomial theorem (1.21) to get

$$\begin{aligned} \exp(x+y) &= \sum_{s=0}^{\infty} \frac{(x+y)^s}{s!} = \sum_{s=0}^{\infty} \frac{1}{s!} \sum_{n=0}^{s} \binom{s}{n} x^n y^{s-n} \\ &= \sum_{s=0}^{\infty} \frac{1}{s!} \sum_{n=0}^{s} \frac{s!}{n!\,(s-n)!} x^n y^{s-n} = \sum_{s=0}^{\infty} \sum_{n=0}^{s} \frac{x^n}{n!} \frac{y^{s-n}}{(s-n)!} \\ &= \sum_{s=0}^{\infty} \sum_{\substack{m\geq 0, n\geq 0 \\ m+n=s}} \frac{x^n}{n!} \frac{y^m}{m!}. \end{aligned}$$

It follows from (2.240) that

$$\exp(x+y) = \lim_{N\to\infty} \sum_{n=0}^{N} \sum_{m=0}^{N} \frac{x^n}{n!} \frac{y^m}{m!} = \lim_{N\to\infty} \sum_{n=0}^{N} \frac{x^n}{n!} \sum_{m=0}^{N} \frac{y^m}{m!} = \exp(x)\exp(y). \tag{2.242}$$

With $x = y$, (2.242) obviously implies $\exp(2x) = [\exp(x)]^2$. That $[\exp(x)]^n = \exp(nx)$ follows from having confirmed the $n = 1$ and $n = 2$ cases, and use of induction: Assuming the result for n, $[\exp(x)]^{n+1} = [\exp(x)]^n \exp(x) = \exp(nx)\exp(x)$, and (2.242) implies $[\exp(x)]^{n+1} = \exp((n+1)x)$. The start of a direct proof of this, using the multinomial theorem, was given in Example 1.14. ■

2.5.5 Sequences and Series of Functions

> It is true that a mathematician who is not also something of a poet will never be a perfect mathematician. (Karl Weierstrass)

Up to this point, f_k represented a series. Now let $\{f_n(x)\}$ be a *sequence of functions* with the same domain, say D. The function f is the *pointwise limit* of sequence $\{f_n\}$, or $\{f_n\}$ *converges pointwise* to f, if, $\forall x \in D$, $\lim_{n\to\infty} f_n(x) = f(x)$. That is, $\forall x \in D$ and for every given $\epsilon > 0$, $\exists N \in \mathbb{N}$ such that $|f_n(x) - f(x)| < \epsilon$, $\forall n > N$. We will also denote this by writing $f_n \to f$. It is helpful to read $\forall x$ in the previous sentence as "for each", and not "for all", to emphasize that N depends on both x and ϵ.

Just as (infinite) series can be associated with sequences, consider the (infinite) *series (of functions)*, $\sum_{n=0}^{\infty} f_n(x)$, $x \in D$. If, for each $x \in D$, the sequence $s_n = \sum_{n=0}^{n} f_n(x)$ converges pointwise, then the series is said to *converge pointwise (on D) to (the function)* $S = \sum_{n=0}^{\infty} f_n(x)$, $x \in D$.

[19] See Example 2.75 and results (2.272) and (2.279) for the justification of termwise differentiation.

Example 2.74 *(Stoll, 2001, p. 320) Let* $f_k(x) = x^2(1+x^2)^{-k}$*, for* $x \in \mathbb{R}$ *and* $k = 0, 1, \ldots$*, and observe that* $f_k(x)$ *is continuous. The series*

$$S(x) := \sum_{k=0}^{\infty} f_k(x) = x^2 \sum_{k=0}^{\infty} \frac{1}{(1+x^2)^k} = 1 + x^2$$

for $x \neq 0$*, and* $S(0) = 0$*. Thus,* $S(x)$ *converges pointwise on* $\mathbb{R}$ *to the function* $f(x) = (1+x^2)\mathbb{I}(x \neq 0)$*, and is not continuous at zero.* ■

The above example shows that the pointwise limit may not be continuous even if each element in the sequence is continuous. Similarly, differentiability or integrability of f_n does not ensure that the pointwise limit shares that property; a standard example for the latter is

$$f_n(x) = nx(1-x^2)^n, \quad x \in [0,1], \tag{2.243}$$

with $\lim_{n\to\infty} f_n(x) = 0$ for $x \in (0,1)$ and $f_n(0) = f_n(1) = 0$, so that the pointwise limit is $f(x) = 0$, but $\int_0^1 f_n \neq \int_0^1 f$ (see, e.g., Stoll, 2001, p. 321; Estep, 2002, §33.3).

The function f is the *uniform limit* on D of the sequence $\{f_n\}$, (or $\{f_n\}$ is *uniformly convergent* to f), if, for every given $\epsilon > 0$, $\exists N \in \mathbb{N}$ such that $|f_n(x) - f(x)| < \epsilon$ for every $x \in D$ whenever $n > N$. The difference between pointwise and uniform convergence parallels that between continuity and uniform continuity; in the uniform limit, N is a function of ϵ, but not of x. In this case, we write $f_n \rightrightarrows f$. Sequence $\{f_n\}$ is said to be *uniformly Cauchy* if, for every given $\epsilon > 0$, $\exists N \in \mathbb{N}$ such that $|f_n(x) - f_m(x)| < \epsilon$ for every $n, m > N$ and every $x \in D$. Let $\{f_n\}$ be a sequence of functions. Important results include:

$$\textbf{If } f_n \rightrightarrows f\textbf{, then } f_n \to f.$$

$$\{f_n\} \textbf{ is uniformly convergent} \iff \{f_n\} \textbf{ is uniformly Cauchy.} \tag{2.244}$$

$$\textbf{If } f_n \in \mathcal{C}^0(D)\textbf{, } \forall n \in \mathbb{N}\textbf{, and } f_n \rightrightarrows f\textbf{, then } f \in \mathcal{C}^0(D). \tag{2.245}$$

$$\textbf{If } f, f_n \in \mathcal{C}^0[a,b]\textbf{, } \forall n \in \mathbb{N}\textbf{, where } [a,b] \subset D \textbf{ and } m_n = \max_x |f_n(x) - f(x)|\textbf{, then}$$
$$f_n(x) \rightrightarrows f(x) \textbf{ on } [a,b] \iff \lim_{n\to\infty} m_n = 0\textbf{.} \tag{2.246}$$

To prove (2.244), suppose $f_n \to f$ uniformly on E. Let $\epsilon > 0$ be given. Then there exists $n_0 \in \mathbb{N}$ such that

$$n \geq n_0 \text{ and } x \in E \Longrightarrow |f_n(x) - f(x)| < \frac{\epsilon}{2}.$$

Hence for all $m, n \geq n_0$ and all $x \in E$, we obtain

$$|f_m(x) - f_n(x)| \leq |f_m(x) - f(x)| + |f(x) - f_n(x)| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Conversely, suppose (f_n) is uniformly Cauchy on E. Then $(f_n(x))$ is a Cauchy sequence in $\mathbb{R}$ for each $x \in E$. For $x \in E$, $(f_n(x))$ converges to a real number, which we denote by $f(x)$. Let $\epsilon > 0$ be given. Then there exists $n_0 \in \mathbb{N}$ such that

$$m, n \geq n_0 \text{ and } x \in E \Longrightarrow |f_m(x) - f_n(x)| < \epsilon.$$

Fix $n \geq n_0$ and let $m \to \infty$. Thus we obtain $|f(x) - f_n(x)| \leq \epsilon$ for all $x \in E$. This implies that $f_n \to f$ uniformly on E.

To prove (2.246), first assume $f_n(x) \rightrightarrows f(x)$ so that, by definition, $\forall \epsilon > 0$, $\exists N \in \mathbb{N}$ such that, $\forall x \in [a,b]$ and $\forall n > N$, $D_n(x) := |f_n(x) - f(x)| < \epsilon$. By assumption, f_n and f are continuous on $[a,b]$, so that $D_n(x)$ is as well, and, from (2.29), $\exists x_n \in [a,b]$ such that $x_n = \arg\max_x D_n(x)$. Thus, $\forall n > N$, $m_n = D_n(x_n) = \max_x D_n(x) < \epsilon$, i.e., $\lim_{n\to\infty} m_n = 0$. Now assume $m_n = \max_x D_n(x) \to 0$ as $n \to \infty$. Then, $\forall \epsilon > 0$, $\exists N \in \mathbb{N}$ such that, $\forall n > N$, $m_n < \epsilon$, so that, $\forall x \in [a,b]$ and $n > N$, $D_n(x) = |f_n(x) - f(x)| \leq m_n < \epsilon$.

The sequence $\{f_n\}$ is said to be *monotone decreasing (increasing)* if, $\forall n \in \mathbb{N}$ and $\forall x \in D$, $f_{n+1}(x) \leq f_n(x)$ ($f_{n+1}(x) \geq f_n(x)$), and *monotone* if it is either monotone decreasing or monotone increasing. The monotonicity of f_n is key for pointwise convergence to imply uniform convergence:

If (i) the $f_n : D \to \mathbb{R}$ are continuous, (ii) the f_n are monotone, (iii) D is a closed, bounded interval, and (iv) $f_n \to f$ to $f \in \mathcal{C}^0$, then

$$f_n \rightrightarrows f \text{ on } D. \tag{2.247}$$

This is often referred to as Dini's Theorem. For proof, see, e.g., Browder (1996, p. 64), Stoll (2021, p. 355), or Ghorpade and Limaye (2018, p. 432).

Example 2.75 *Let $f_n(x) = x^n/n!$, $n = 0, 1, \ldots$ and $s_n(x) = \sum_{k=0}^n f_n(x)$. Then (i) $f_n(x), s_n(x) \in \mathcal{C}^0$ for $x \in \mathbb{R}$ and all $n \in \mathbb{N}$, (ii) for each $x \geq 0$, $s_n(x)$ is monotone increasing in n, (iii) $\forall r \in \mathbb{R}_{>0}$, $D = [0,r]$ is a closed, bounded interval, and (iv) from Example 2.73, $s_n \to \exp(x)$ for all $x \in \mathbb{R}$. Thus, from Dini's theorem (2.247), $\forall r \in \mathbb{R}_{>0}$ and $\forall x \in [0,r]$, $\lim_{n\to\infty} s_n(x) = \sum_{k=0}^\infty x^n/n! \rightrightarrows \exp(x)$.* ■

Example 2.76 *Let $E := [-1,1]$ and $f_n(x) := \sqrt{x^2 + (1/n^2)}$ for $n \in \mathbb{N}$ and $x \in E$. Then $f_n \to f$ on E, where $f(x) := \sqrt{x^2} = |x|$ for $x \in E$. Note that $[-1,1]$ is a closed and bounded subset of $\mathbb{R}$, $f_n \geq f_{n+1}$ for all $n \in \mathbb{N}$ on E, and each f_n is continuous on E, and so is f. Hence by Dini's theorem (2.247), $f_n \to f$ uniformly on E.* ■

Example 2.77 *(Ghorpade and Limaye, p. 433) We give examples to show that none of the hypotheses in the Dini Theorem for sequences of continuous functions can be omitted.*

(a) Let $E := [0,2]$, and for each $n \in \mathbb{N}$, let $f_n(x) := 1 - |2nx - 3|$ if $(1/n) \leq x \leq (2/n)$ and $f_n(x) := 0$ otherwise. In this example, the set E is closed and bounded, but the sequence (f_n) is not monotonic.

(b) Let $E := (0,1]$, and for each $n \in \mathbb{N}$, let $f_n(x) := 1/(nx+1)$ for $x \in E$. In this example, the sequence (f_n) is monotonic and the set E is bounded but E is not closed.

(c) Let $E := [1,\infty)$, and for each $n \in E$, let $f_n(x) := x/(x+n)$ for $x \in E$. In this example, the sequence (f_n) is monotonic and the set E is closed, but E is not bounded.

(d) Let $E := [0,1]$, and for $n \in E$, let $f_n(x) := 1 - nx$ if $0 \leq x \leq (1/n)$ and $f_n(x) := 0$ otherwise. Here the set E is closed and bounded, the sequence (f_n) is monotonic, and it converges to a discontinuous function $f : E \to \mathbb{R}$ given by $f(0) := 1$ and $f(x) := 0$ if $x \in (0,1]$.

In (a), (b), and (c) above, $f_n \to f$ on E, where $f := 0$ on E, but the sequence (f_n) does not converge to f uniformly, since $\sup\{|f_n(x) - f(x)| : x \in E\} = 1$ for each $n \in \mathbb{N}$. In (d) above, (f_n) does not converge uniformly to f, since $|f_n(1/2n) - f(1/2n)| = 1/2$ for each $n \in \mathbb{N}$. ■

Before discussing series (sums of functions), we take a short detour to discuss another form of convergence of functions, namely mean square. We will not make subsequent use of this, but it is of paramount importance in, e.g., Fourier (more generally, harmonic) analysis, where clear, useful, convenient results can be derived using mean square convergence, but far less so for pointwise and uniform convergence. We restrict attention to continuous functions, which, thus, are Riemann integrable.

Definition: Denote by $C(I)$ the space of continuous functions on $I = [a, b]$.

(a) The mean square norm, or L^2 norm, denoted $\|f\|_2$, of $f \in C(I)$ is given by

$$\|f\|_2 = \left(\int_I |f(x)|^2 dx \right)^{1/2}.$$

(b) By the mean square distance, or simply the distance, between functions $f, g \in C(I)$, we mean the quantity

$$\|f - g\|_2 = \left(\int_I |f(x) - g(x)|^2 dx \right)^{1/2}.$$

(c) If $f, f_1, f_2, \ldots \in C(I)$ satisfy

$$\lim_{N \to \infty} \|f_N - f\|_2 = 0,$$

then we say the sequence $f_1, f_2, \ldots$ converges to f, or the f_N 's converge to f, in mean square norm.

More generally,

Definition: Given $1 \leq p \leq \infty$, for each $f \in C[a, b]$,

$$\|f\|_p = \begin{cases} \left(\int_a^b |f(t)|^p dt \right)^{1/p}, & \text{if } 1 \leq p < \infty, \\ \sup_{t \in [a,b]} |f(t)|, & \text{if } p = \infty. \end{cases}$$

It is shown in books on measure theory, metric space analysis, and functional analysis, that $\| \cdot \|_p$ is a norm on $C[a, b]$. We call $\| \cdot \|_p$ the L^p-norm on $C[a, b]$.

Theorem: Let $I = [a, b]$, let $f_1, f_2, \ldots$ be a sequence of functions in $C(I)$; and let $f \in C(I)$. If this sequence converges uniformly to f, then it converges in norm to f.

Proof: By basic properties of the Riemann integral,

$$\begin{aligned} \|f_N - f\|^2 &= \int_I |f_N(x) - f(x)|^2 \, dx \leq \sup_{x \in I} |f_N(x) - f(x)|^2 \int_I dx \\ &= (b - a) \left(\sup_{x \in I} |f_N(x) - f(x)| \right)^2, \end{aligned} \tag{2.248}$$

the last step because "the sup of the square equals the square of the sup", from (2.202).

If the f_N's converge uniformly to f, then the right side of the above goes to zero as $N \to \infty$. By the Squeeze theorem, then, so does $\|f_N - f\|^2$, and therefore so does $\|f_N - f\|$. So the f_N's converge in norm to f, as required.

We return to the function used in Example 2.13 to show that pointwise convergence does not imply convergence in L^2 norm.

Example 2.78 *(Stade, Fourier Analysis, p. 158) For domain $D = [0, 2\pi]$, $N \in \mathbb{N}$, let $f_N : D \to \mathbb{R}$ be the function defined by $f_N(x) = N\left(\frac{x}{2\pi}\right)^N \sqrt{2\pi - x}$. Also let $f(x) = 0$ for all $x \in [0, 2\pi]$. Then, with substitution $x = 2\pi u$,*

$$\begin{aligned}
\lim_{N\to\infty} \|f_N - f\|^2 &= \lim_{N\to\infty} \int_0^{2\pi} |f_N(x) - f(x)|^2 \, dx \\
&= \lim_{N\to\infty} N^2 \int_0^{2\pi} \left(\frac{x}{2\pi}\right)^{2N} (2\pi - x) dx \\
&= 4\pi^2 \lim_{N\to\infty} N^2 \int_0^1 u^{2N}(1-u) dx \\
&= 4\pi^2 \lim_{N\to\infty} N^2 \left[\frac{u^{2N+1}}{2N+1} - \frac{u^{2N+2}}{2N+2}\right]_0^1 \\
&= 4\pi^2 \lim_{N\to\infty} N^2 \frac{1}{(2N+1)(2N+2)} = \pi^2 \neq 0.
\end{aligned}$$

Thus, $\|f_N - f\| \nrightarrow 0$, so the f_N's do not converge to f in norm. Notice that $\|f_N - f\|$ does converge to something. ■

Let $\{f_n\}$ be a sequence of functions, each with domain D. The series $\sum_{n=1}^{\infty} f_n(x)$ is said to *converge uniformly (on D) to the function S* if the associated sequence of partial sums converges uniformly on D. In this case, we write $\sum_{n=1}^{\infty} f_n(x) \rightrightarrows S(x)$. For sequence $\{f_n(x)\}$ with domain D:

$$\textbf{If, } \forall n \in \mathbb{N},\ f_n \in \mathcal{C}^0 \textbf{ and } \sum_{n=1}^{\infty} f_n(x) \rightrightarrows S(x), \textbf{ then } S \in \mathcal{C}^0. \tag{2.249}$$

For proof, see, e.g., Stoll (2021, p. 353).

The next result is the famous *Weierstrass M-test*, after Karl Theodor Wilhelm Weierstrass (1815–1897):

> **If there exists a sequence of constants M_n such that, $\forall x \in D$ and $\forall n \in \mathbb{N}$, $|f_n(x)| \leq M_n$, and $\sum_{n=1}^{\infty} M_n < \infty$, then $\sum_{n=1}^{\infty} f_n(x)$ is uniformly convergent on D.**

Proof: Let $S_n(x) = \sum_{k=1}^{n} f_k(x)$ be the nth partial sum of sequence f_k. Then, for $n > m$,

$$|S_n(x) - S_m(x)| = \left|\sum_{k=m+1}^{n} f_k(x)\right| \leq \sum_{k=m+1}^{n} |f_k(x)| \leq \sum_{k=m+1}^{n} M_k. \tag{2.250}$$

As the series M_n is convergent, the Cauchy criterion (2.213) implies that the rhs of (2.250) can be made arbitrarily close to zero. The result now follows from (2.244).

Example 2.79 *Let* $f_n(x) = (-1)^n x^{2n}/(2n)!$, $x \in \mathbb{R}$. *Then,* $\forall L \in \mathbb{R}_{>0}$,

$$|f_n(x)| = \left|\frac{x^{2n}}{(2n)!}\right| \le \left|\frac{L^{2n}}{(2n)!}\right| =: M_n, \quad x \in D = [-L, L].$$

By the ratio test, $M_{n+1}/M_n = L^2/(2n+1)(2n+2) \to 0$ *as* $n \to \infty$, *so that* $\sum_{n=0}^{\infty} M_n < \infty$, *and, from the Weierstrass M-test,* $\sum_{n=0}^{\infty} f_n(x)$ *converges uniformly on* $D = [-L, L]$. *This justifies the definitions*

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

as in (2.58).[20] *Next, for* $x \neq 0$, *we wish to know if*

$$\frac{\sin x}{x} = \frac{1}{x} \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!} \stackrel{?}{=} \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n+1)!}.$$

As the series for $\sin x$ *is convergent, it follows from the definition of convergence that, for any* $\epsilon > 0$ *and* $x \neq 0$, $\exists N = N(x, \epsilon) \in \mathbb{N}$ *such that,* $\forall n > N$,

$$\left|\sin x - \sum_{n=0}^{N} \frac{(-1)^n x^{2n+1}}{(2n+1)!}\right| < \epsilon |x| \quad \text{or} \quad \left|\frac{\sin x}{x} - \sum_{n=0}^{N} \frac{1}{x} \frac{(-1)^n x^{2n+1}}{(2n+1)!}\right| < \epsilon,$$

so that, for $x \neq 0$,

$$\frac{\sin x}{x} = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n+1)!}. \tag{2.251}$$

With $0^0 = 1$, *the rhs equals 1 for* $x = 0$, *which coincides with the limit as* $x \to 0$ *of the lhs. As above, the Weierstrass M-test shows that this series is uniformly convergent on* $[-L, L]$ *for any* $L \in \mathbb{R}_{>0}$. ■

Example 2.80 *(Example 2.75 cont.) Again let* $f_n(x) = x^n/n!$, $n = 0, 1, \ldots$, *and* $s_n(x) = \sum_{k=0}^{n} f_n(x)$. *For all* $L \in \mathbb{R}_{>0}$,

$$|f_n(x)| = \frac{|x^n|}{n!} \le \frac{L^n}{n!} = M_n, \quad x \in [-L, L],$$

and $\sum_{n=0}^{\infty} M_n < \infty$ *from Example 2.73, which showed that* $s_n(L) \to \exp(L)$ *absolutely. Thus, the Weierstrass M-test implies that* $\sum_{k=0}^{\infty} f_n(x)$ *converges uniformly on* $[-L, L]$, *where* L *is an arbitrary positive real number. It is, however,* not *true that* $\sum_{k=0}^{\infty} x^n/n!$ *converges uniformly on* $(-\infty, \infty)$. *Also, as required below,*

$$\sum_{r=0}^{\infty} \frac{(-z \ln z)^r}{r!} \rightrightarrows e^{-z \ln z}, \tag{2.252}$$

by taking $x = -z \ln z \in \mathbb{R}$. ■

[20] Observe that, while $\lim_{n\to\infty} x^{2n}/(2n)! = 0$ for all $x \in \mathbb{R}$, for any fixed n, $\lim_{x\to\infty} x^{2n}/(2n)! = \infty$. This means that, although the series converges, evaluation of the truncated sum will be numerically problematic because of the limited precision with which numbers are digitally stored. Of course, the relations $\cos(x+\pi) = -\cos x$, $\cos(x+2\pi) = \cos x$ and $\cos(-x) = \cos x$; and $\sin(x+\pi) = -\sin x$, $\sin(x+2\pi) = \sin x$ and $\sin(-x) = -\sin x$; and $\sin x = -\cos(x + \pi/2)$, imply that only the series for cosine is required, with x restricted to $[0, \pi/2]$. For $x \in [0, \pi/2]$, it is enough to sum the $\cos x$ series up to $n = 10$ to ensure 15 digit accuracy.

Example 2.81 *Let* $f_k(x) = (x/R)^k =: a^k$ *for a fixed* $R \in \mathbb{R}_{>0}$. *Then* $G(x) = \sum_{k=1}^{\infty} f_k(x)$ *is a geometric series that converges to* $S(x) = (x/R)/(1 - x/R)$ *for* $|x/R| < 1$, *or* $|x| < R$. *Let* S_n *be the partial sum of sequence* $\{f_k\}$. *For* $G(x)$ *to converge uniformly to* $S(x)$, *it must be the case that, for any* $\epsilon > 0$, *there is a value* $N \in \mathbb{N}$ *such that,* $\forall n > N$,

$$|S_n - G| = \left| \sum_{k=1}^{n} \left(\frac{x}{R}\right)^k - \frac{x/R}{1 - x/R} \right| = \left| \frac{a - a^{n+1}}{1 - a} - \frac{a}{1 - a} \right| = \left| \frac{a^{n+1}}{1 - a} \right| < \epsilon. \tag{2.253}$$

But, for any n,

$$\lim_{x \to R^-} \frac{a^{n+1}}{1 - a} = \lim_{a \to 1^-} \frac{a^{n+1}}{1 - a} = \infty,$$

so that the inequality in (2.253) cannot hold. Now choose a value b *such that* $0 < b < R$ *and let* $M_k = (b/R)^k$, *so that* $\sum_{k=1}^{\infty} M_k = (b/R)/(1 - b/R) < \infty$. *Then, for* $|x| \leq b$, $\left|(x/R)^k\right| \leq (b/R)^k = M_k$, *and use of the Weierstrass M-test shows that the series* $G(x)$ *converges uniformly on* $[-b, b]$ *to* $S(x)$. *See also Example 2.92.* ■

Remarks:

(a) When using the Maple engine that accompanies Scientific Workplace 4.0, evaluating $\lim_{n\to\infty} \sum_{k=1}^{n} (-1)^k$ yields the interval $-1..0$, yet evaluating $\sum_{k=1}^{\infty} (-1)^k$ produces $-1/2$. Presumably, the latter result is obtained because Maple computes $\lim_{x\to -1^+} \sum_{k=1}^{\infty} x^k = \lim_{x\to -1^+} x/(1 - x) = -1/2$, which is itself correct, but,

$$\sum_{k=1}^{\infty} (-1)^k \neq \lim_{x \to -1^+} \sum_{k=1}^{\infty} x^k.$$

From (2.249), this would be true if $\sum_{k=1}^{\infty} x^k$ were uniformly convergent for $x = -1$, which it is not. While this is probably a mistake in Maple, it need not be one, as the next remark shows.

(b) The series $\sum_{k=1}^{\infty} a_k$ is said to be *Abel summable* to L if $\lim_{x\to 1^-} f(x) = L$, where $f(x) = \sum_{k=1}^{\infty} a_k x^k$ for $0 \leq x < 1$ (after Neils Henrik Abel, 1802–1829; see Goldberg, 1964, p. 251). For example, with $a_k = (-1)^k$, the series $f(x) = \sum_{k=1}^{\infty} a_k x^k = -x + x^2 - x^3 + \cdots$ converges for $|x| < 1$ to $-x/(x + 1)$. Then the series $-1 + 1 - 1 + 1 - \cdots$ is clearly divergent, but is Abel summable to $\lim_{x\to 1^-} f(x) = -1/2$. ■

Example 2.82 *The contrapositive of (2.249) implies*

$$f_n \in \mathcal{C}^0,\ f \notin \mathcal{C}^0 \implies \sum_{n=1}^{\infty} f_n(x) \not\rightrightarrows f.$$

In words, if the f_n *are continuous but* f *is not, then* $\sum_{n=1}^{\infty} f_n(x)$ *is not uniformly convergent to* f. *In Example 2.74,* $f(x)$ *is not continuous at* $x = 0$, *so that* $\sum_{k=0}^{\infty} f_k(x)$ *is not uniformly convergent on any interval containing zero.* ■

Recall from (2.245) and (2.249) that uniform convergence of sequences and series of continuous functions implies continuity of the limiting function. Similar result hold for integrability of sequences and series:

If, $\forall n \in \mathbb{N}$, $f_n \in \mathcal{R}[a, b]$ **and** $f_n(x) \rightrightarrows f(x)$ **on** $[a, b]$, **then** $f \in \mathcal{R}[a, b]$ **and**

$$\lim_{n\to\infty}\int_a^b f_n(x)\,dx = \int_a^b \lim_{n\to\infty} f_n(x)\,dx = \int_a^b f(x)\,dx. \tag{2.254}$$

The proof is worth showing, as it demonstrates why the uniform convergence assumption is necessary.

Proof: Let $\epsilon_n = \max_{x\in[a,b]} |f_n(x) - f(x)|$. Uniform convergence implies that $\lim_{n\to\infty}\epsilon_n = 0$ using (2.246), and that

$$f_n(x) - \epsilon_n \le f(x) \le f_n(x) + \epsilon_n. \tag{2.255}$$

Recall the definitions of $\underline{S}(f,\pi)$ and $\overline{S}(f,\pi)$ from §2.4.1. As $f_n \in \mathcal{R}[a,b]$,

$$\sup_\pi \underline{S}(f_n,\pi) = \inf_\pi \overline{S}(f_n,\pi) = \int_a^b f_n(x)\,dx.$$

Further, $f_n(x) - \epsilon_n \in \mathcal{R}[a,b]$, so

$$\sup_\pi \underline{S}(f_n-\epsilon_n,\pi) = \inf_\pi \overline{S}(f_n-\epsilon_n,\pi) = \int_a^b (f_n(x)-\epsilon_n)dx$$

and, from (2.255),

$$\underline{S}(f_n-\epsilon_n,\pi) < \underline{S}(f,\pi), \quad \text{and} \quad \overline{S}(f,\pi) < \overline{S}(f_n+\epsilon_n,\pi).$$

Thus, $\forall n \in \mathbb{N}$,

$$\int_a^b (f_n(x)-\epsilon_n)\,dx \le \sup_\pi \underline{S}(f,\pi) \le \inf_\pi \overline{S}(f,\pi) \le \int_a^b (f_n(x)+\epsilon_n)\,dx. \tag{2.256}$$

This implies

$$0 \le \inf_\pi \overline{S}(f,\pi) - \sup_\pi \underline{S}(f,\pi) \le \int_a^b (f_n(x)+\epsilon_n)\,dx - \int_a^b (f_n(x)-\epsilon_n)\,dx = \int_a^b 2\epsilon_n dx,$$

but $\lim_{n\to\infty}\epsilon_n = 0$, so that $\lim_{n\to\infty}\left[\inf_\pi \overline{S}(f,\pi) - \sup_\pi \underline{S}(f,\pi)\right] = 0$, and $f \in \mathcal{R}[a,b]$. Now we can write

$$\begin{aligned}\left|\int_a^b f_n(x)\,dx - \int_a^b f(x)\,dx\right| &= \left|\int_a^b [f_n(x)-f(x)]\,dx\right| \\ &\le \int_a^b |f_n(x)-f(x)|\,dx \le \int_a^b \epsilon_n\,dx = \epsilon_n(b-a),\end{aligned}$$

and taking the limit yields (2.254).

If, $\forall n \in \mathbb{N}$, $f_n \in \mathcal{R}[a,b]$ **and** $\sum_{n=1}^\infty f_n(x) \rightrightarrows S(x)$ **for** $x \in [a,b]$**, then** $S \in \mathcal{R}[a,b]$ **and**

$$\int_a^b \left(\lim_{n\to\infty}\sum_{k=1}^n f_k(x)\right)dx = \int_a^b S(x)\,dx = \sum_{n=1}^\infty \int_a^b f_n(x)\,dx = \lim_{n\to\infty}\sum_{k=1}^n \int_a^b f_k(x)\,dx,$$

i.e.,

$$\int_a^b S(x)\,dx = \sum_{n=1}^{\infty} \int_a^b f_n(x)\,dx. \tag{2.257}$$

Proof: et $S_n(x) = \sum_{k=1}^{n} f_k(x)$, so that, $\forall n \in \mathbb{N}$, $S_n \in \mathcal{R}[a,b]$. From the previous result, $S \in \mathcal{R}[a,b]$ and, from (2.254) applied to $S(x)$ and $S_n(x)$,

$$\int_a^b \sum_{k=1}^{\infty} f_k(x)\,dx = \int_a^b S(x)\,dx = \lim_{n\to\infty} \int_a^b S_n(x)\,dx = \lim_{n\to\infty} \sum_{k=1}^{n} \int_a^b f_k(x)\,dx, \tag{2.258}$$

which is (2.257), having used finite additivity (2.134) of the Riemann integral.

Example 2.83 *Recall from Example 2.9 that* $\lim_{x\to 0} x^x = 1$. *The integral* $I = \int_0^1 x^{-x} dx$ *was shown to be equal to* $\sum_{r=1}^{\infty} r^{-r}$ *by Johann Bernoulli in 1697. To see this, as in Havil (2003, p. 44), use (2.90) and (2.241) to write*

$$I = \int_0^1 e^{-x\ln x} dx = \int_0^1 \sum_{r=0}^{\infty} \frac{(-x\ln x)^r}{r!} dx = \sum_{r=0}^{\infty} \frac{(-1)^r}{r!} \int_0^1 (x\ln x)^r\,dx,$$

where the exchange of sum and integral is justified by (2.252) and (2.257). The result now follows from (2.161), i.e., $\int_0^1 (x\ln x)^r\,dx = (-1)^r\, r!\ /\ (r+1)^{r+1}$, *or*

$$I = \sum_{r=0}^{\infty} \frac{(-1)^r}{r!} \frac{(-1)^r\, r!}{(r+1)^{r+1}} = \sum_{r=0}^{\infty} \frac{1}{(r+1)^{r+1}}. \qquad \blacksquare$$

Example 2.84 *(Browder, 1996, p. 113) Let* $f_n(x) = x/[n(x+n)]$ *for* $x \in I = [0,1]$ *and* $n \in \mathbb{N}$. *As* $f_n'(x) = 1/(x+n^2) > 0$ *for* $x \in I$, $\max f_n$ *occurs at* $x = 1$. *Thus,* $0 \le f_n(x) \le 1/n(n+1)$, *and, from the comparison test with* $g_n = n^{-2}$, $\sum_{n=1}^{\infty} [n(n+1)]^{-1}$ *converges. This series has the "telescoping property", i.e.,*

$$\frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{n+1}, \qquad \textit{so that} \qquad \sum_{k=1}^{n} \frac{1}{k(k+1)} = \frac{n}{n+1} \to 1. \tag{2.259}$$

Thus, from the Weierstrass M-test, $\sum_{n=1}^{\infty} f_n(x)$ *converges uniformly on* $[0,1]$ *to a function, say* $S(x)$, *which, by (2.249), is continuous. Further, from (2.259),* $0 \le S(x) \le 1$. *From (2.257),*

$$\sum_{n=1}^{\infty} \int_0^1 f_n(x)\,dx = \int_0^1 S(x)\,dx =: \gamma, \tag{2.260}$$

and $\int_0^1 S(x)\,dx < \int_0^1 dx = 1$, *so that* $0 < \gamma < 1$. *From (2.155) (or perform the substitution* $u = x+n$*),* $\int_0^1 1/(x+n)dx = \ln(n+1) - \ln n$, *so*

$$\int_0^1 f_n = \int_0^1 \left(\frac{1}{n} - \frac{1}{x+n}\right) dx = \frac{1}{n} - \ln\frac{n+1}{n},$$

and (2.260) implies

$$\gamma = \lim_{N\to\infty} \sum_{n=1}^{N} \int_0^1 f_n(x)\,dx = \lim_{N\to\infty} \left[\sum\nolimits_{n=1}^{N} n^{-1} - \ln(N+1)\right],$$

or, as $\lim_{N\to\infty}[\ln(N+1)-\ln N]=0$,

$$\gamma=\lim_{N\to\infty}\left(\sum_{n=1}^{N}\frac{1}{n}-\ln N\right),$$

which is Euler's constant from Example 2.64. ■

Example 2.85 *Let* $S(x)=1-x+x^2-x^3+\cdots$. *For* $-1<x\le 0$, $S(x)=1+y+y^2+\cdots$, *where* $y=-x$, *and so converges to* $1/(1-y)=1/(1+x)$ *from (2.215). For* $0\le x<1$, *the alternating series test (page 105) shows that* $S(x)$ *converges; and from (2.216), converges to* $1/(1+x)$. *Thus,* $S(x)$ *converges to* $1/(1+x)$ *for* $|x|<1$.

Similar to the derivation in Example 2.81, for every $b\in[0,1)$, $S(x)$ *is uniformly convergent for* $x\in[-b,b]$. *So, from (2.257),*

$$\int_0^b\frac{1}{1+x}\,dx=\int_0^b 1\,dx-\int_0^b x\,dx+\int_0^b x^2\,dx-\cdots.$$

For the first integral, let $u=1+x$ *so that* $\int_0^b(1+x)^{-1}\,dx=\int_1^{b+1}u^{-1}\,du=\ln(1+b)$. *Thus,*

$$\ln(1+b)=b-\frac{b^2}{2}+\frac{b^3}{3}-\cdots,\qquad |b|<1. \tag{2.261}$$

Example 2.95 will show that (2.261) also holds for $b=1$. ■

Example 2.86 *Similar to Example 2.85,*

$$\frac{1}{1+y^2}=1-y^2+y^4-y^6+\cdots,\quad |y|<1, \tag{2.262}$$

and, for every $b\in[0,1)$, *the rhs is uniformly convergent for* $y\in[-b,b]$, *so that termwise integration of the rhs is permitted. From (2.61) and the FTC (2.143),*

$$\int_0^t\frac{1}{1+y^2}\,dy=\arctan(t)-\arctan(0)=\arctan(t), \tag{2.263}$$

so that

$$\arctan(t)=t-\frac{t^3}{3}+\frac{t^5}{5}-\cdots=\sum_{n=0}^{\infty}\frac{(-1)^n}{2n+1}t^{2n+1},\quad t\in[0,1), \tag{2.264}$$

from which $\arctan(t)$, $t\in[0,1)$, *can be computed to any degree of accuracy up to machine precision.*[21] *Example 2.96 below consider the case when* $t=1$. ■

Another useful result is the *bounded convergence theorem*, which involves the interchange of limit and integral using only pointwise convergence, but requires that f also be integrable on $I=[a,b]$, and that the f_n are bounded for all n and all $x\in I$.

> **If,** $\forall n\in\mathbb{N}$, $f_n\in\mathcal{R}[a,b]$ **with** $f_n(x)\to f(x)$, $f\in\mathcal{R}[a,b]$, **and** $\exists M\in\mathbb{R}_{>0}$ **such that,** $\forall x\in[a,b]$ **and** $\forall n\in\mathbb{N}$, $|f_n(x)|\le M$, **then**

[21] One could use so-called virtual precision arithmetic, or VPA, to obtain far higher accuracy. It works by using software, and traditional computing machines that allocate 64 bits per number, to allow computations with any (up to some limit) precision.

$$\lim_{n\to\infty}\int_a^b f_n(x)\,dx = \int_a^b f(x)\,dx. \tag{2.265}$$

This is explicitly proven in Stoll (2001, §10.6) and is a special case of *Lebesgue's dominated convergence theorem*, detailed in, e.g., Browder (1996, §10.2), Stoll (2001, §10.7), Pugh (2002, §6.4), and any book on measure theory and the Lebesgue integral. Interestingly, measure theory is not required to prove it. The Arzela Bounded Convergence Theorem proves the result using only basic analysis, and can be found in, e.g., Ghorpade and Limaye (2018, Prop. 10.40).

Paralleling results (2.254) and (2.257), let $\{f_n(x)\}$ be a sequence of functions with nth partial sum $S_n(x)$, such that $S_n(x) \to S(x)$. If the conditions leading to (2.265) apply to $\{S_n(x)\}$ and $S(x)$, then (2.257) also holds, the proof of which is the same as (2.258).

Example 2.87 *With $f_n(x) = nx(1-x^2)^n$ as in (2.243), use of a symbolic software package easily shows that $f_n'(x) = -n(1-x^2)^{n-1}(x^2(1+2n)-1)$ and solving $f_n'(x_m) = 0$ yields $x_m = (1+2n)^{-1/2}$, so that*

$$f_n(x_m) = \frac{n}{\sqrt{1+2n}}\left(\frac{2n}{1+2n}\right)^n \quad \text{and} \quad \lim_{n\to\infty} f_n(x_m) = \infty.$$

Thus, $\nexists M$ such that, $\forall x \in [0,1]$ and $\forall n \in N$, $|f_n(x)| \le M$, and the contrapositive of the bounded convergence theorem implies that $\lim_{n\to\infty}\int_0^1 f_n \neq \int_0^1 f$, as determined with a direct calculation of the integrals, as mentioned just after (2.243). ■

Example 2.88 *For a fixed $x \in \mathbb{R}_{>0}$ and all $t \in \mathbb{R}$, define*[22]

$$h(t) := \frac{\exp(-x(1+t^2))}{1+t^2} = e^{-x}\frac{e^{-xt^2}}{1+t^2}, \tag{2.266}$$

which is the integrand in (2.188). Interest centers on developing computational formulae for $\int_0^1 h$ and comparing their efficacy.

Method 1. *From (2.241) with x replaced by $-xt^2$,*

$$h(t) = \frac{e^{-x}}{1+t^2}\left(1 - xt^2 + \frac{x^2t^4}{2!} - \cdots\right) = e^{-x}\sum_{k=0}^{\infty}(-1)^k\frac{x^k}{k!}\frac{t^{2k}}{1+t^2},$$

and termwise integration, valid from Example 2.80 and (2.257), gives

$$\int_0^1 h(t)\;dt = e^{-x}\sum_{k=0}^{\infty}(-1)^k\frac{x^k}{k!}J_k,$$

where

$$J_k := \int_0^1 \frac{t^{2k}}{1+t^2}dt, \quad k \in \mathbb{N},$$

which seems resilient to use of a transformation or integration by parts. It is, however, quite simple, with the following "trick":

$$\int \frac{t^{2k}}{1+t^2}\,dt = \int \frac{t^{2k-2}(t^2+1-1)}{1+t^2}\,dt = \int \frac{t^{2k-2}(1+t^2)}{1+t^2}\,dt - \int \frac{t^{2k-2}}{1+t^2}\,dt,$$

[22] This example was contributed by my friend and colleague Professor Walther Paravicini.

giving the recursion

$$\int_a^b \frac{t^{2k}}{1+t^2}\,dt = \frac{t^{2k-1}}{2k-1}\bigg|_a^b - \int_a^b \frac{t^{2k-2}}{1+t^2}\,dt.$$

With $a = 0$ *and* $b = 1$,

$$J_k = \frac{1}{2k-1} - J_{k-1}. \tag{2.267}$$

From (2.189), $J_0 = \pi/4$, *and iterating (2.267) gives*

$$J_1 = 1 - \pi/4, \quad J_2 = 1/3 - 1 + \pi/4, \quad J_3 = 1/5 - 1/3 + 1 - \pi/4$$

and the general formula

$$J_k = (-1)^k \left(\sum_{m=1}^{k} \frac{(-1)^m}{2m-1} + \pi/4 \right),$$

so that

$$\int_0^1 h\,(t)\;\,dt = e^{-x} \sum_{k=0}^{\infty} \frac{x^k}{k!}\;\left(\sum_{m=1}^{k} \frac{(-1)^m}{2m-1} + \pi/4 \right). \tag{2.268}$$

This sum converges very fast because of the $k!$ *in the denominator. To illustrate, take* $x = 0.3$. *Accurate numeric integration of* h, *and also evaluation of (2.268) truncating the infinite sum at* $U = 200$, *gives* 0.5378448777, *which we will deem correct to 10 digits. With only* $U = 4$, *(2.268) yields* 0.5378456, *accurate to 5 digits.*

Method 2. *We make the* ansatz *that* $h\,(t)$ *can be expressed as the series* $h(t) = \sum_{k=0}^{\infty} a_k t^{2k}$, *and calculate the* a_k. *With* $j = k - 1$,

$$\begin{aligned} e^{-xt^2} &= e^x(1+t^2)\sum_{k=0}^{\infty} a_k t^{2k} = e^x \left(a_0 + \sum_{k=1}^{\infty} a_k t^{2k} + \sum_{k=0}^{\infty} a_k t^{2k+2} \right) \\ &= e^x \left(a_0 + \sum_{j=0}^{\infty} a_{j+1} t^{2j+2} + \sum_{k=0}^{\infty} a_k t^{2k+2} \right) = e^x \left(a_0 + \sum_{j=0}^{\infty} (a_{j+1} + a_j)\, t^{2j+2} \right) \\ &= e^x a_0 + e^x \sum_{k=1}^{\infty} (a_k + a_{k-1})\, t^{2k}. \end{aligned}$$

With $t = 0$, *it follows immediately that* $a_0 = e^{-x}$. *By comparison with* $\exp(-xt^2) = \sum_{k=0}^{\infty} (-xt^2)^k/k!$, *we see that*

$$e^x(a_k + a_{k-1}) = \frac{(-1)^k}{k!} x^k, \qquad k = 1, 2, \ldots .$$

Iterating on

$$e^x a_k = -e^x a_{k-1} + \frac{(-1)^k}{k!} x^k$$

with $a_0 = e^{-x}$ *gives*

$$\begin{aligned} e^x a_1 &= -1 + \frac{(-1)^1}{1!} x^1 = -\left(1 + \frac{x^1}{1!} \right), \\ e^x a_2 &= +\left(1 + \frac{x^1}{1!} \right) + \frac{(-1)^2}{2!} x^2 = +\left(1 + \frac{x^1}{1!} + \frac{x^2}{2!} \right), \end{aligned}$$

and, in general,

$$e^x a_k = (-1)^k \left(1 + \frac{x^1}{1!} + \frac{x^2}{2!} + \cdots + \frac{x^k}{k!}\right) = (-1)^k \sum_{j=0}^{k} \frac{x^j}{j!}.$$

Thus,

$$h(t) = \sum_{k=0}^{\infty} a_k t^{2k} = e^{-x} \sum_{k=0}^{\infty} (-1)^k \left(\sum_{j=0}^{k} \frac{x^j}{j!}\right) t^{2k}$$

and, as $\int_0^1 t^{2k} dt = 1/(2k+1)$,

$$\int_0^1 h(t)\ dt = e^{-x} \sum_{k=0}^{\infty} (-1)^k \frac{1}{2k+1} \left(\sum_{j=0}^{k} \frac{x^j}{j!}\right). \tag{2.269}$$

Whereas (2.268) has a $k!$ *in the denominator, (2.269) has only* $2k+1$*, so we expect it to converge much slower. Indeed, with* $x = 0.3$*, use of* 1000 *terms in the sum results in* 0.53809*, which is correct only to three digits. The formula is useless for numeric purposes.*

Method 3*. Expanding the numerator of the middle term in (2.266) as a power series in* $-x(1+t^2)$ *gives*

$$h(t) = \frac{1}{1+t^2} \sum_{k=0}^{\infty} (-1)^k \frac{x^k}{k!} \left(1+t^2\right)^k = \sum_{k=0}^{\infty} (-1)^k \frac{x^k}{k!} \left(1+t^2\right)^{k-1},$$

so that

$$\int_0^1 h(t)\ dt = \sum_{k=0}^{\infty} (-1)^k \frac{x^k}{k!} I_k,$$

where $I_k := \int_0^1 (1+t^2)^{k-1}\ dt$*. From (2.189),* $I_0 = \pi/4$*, and for* $k > 0$*, use of the binomial formula gives*

$$I_k = \int_0^1 \sum_{m=0}^{k-1} \binom{k-1}{m} t^{2m}\ dt = \sum_{m=0}^{k-1} \binom{k-1}{m} \frac{1}{2m+1},$$

yielding

$$\int_0^1 h(t)\ dt = \frac{\pi}{4} + \sum_{k=1}^{\infty} (-1)^k \frac{x^k}{k!} \sum_{m=0}^{k-1} \binom{k-1}{m} \frac{1}{2m+1}. \tag{2.270}$$

Like (2.268), (2.270) converges fast: with $x = 0.3$*, truncating the infinite sum at* $U = 6$ *gives* 0.5378453*, which is accurate to 5 digits. Based on this value of* x*, it appears that (2.268) converges the fastest.* ■

Example 2.89 *Consider evaluating the improper integral* $\int_0^\infty e^{-sx} x^{-1} \sin x\, dx$ *for* $s \in \mathbb{R}_{>1}$*. From (2.251), (2.5), and that* $x > 0$*,*

$$\frac{\sin x}{x} = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n+1)!} < \sum_{n=0}^{\infty} \frac{x^{2n}}{(2n+1)!} < \sum_{n=0}^{\infty} \frac{x^{2n}}{(2n)!} < \sum_{n=0}^{\infty} \frac{x^n}{(n)!} = e^x,$$

so that, as $x > 0$ *and* $s > 1$*,*

$$e^{-sx} \frac{\sin x}{x} < e^{-sx} e^x = e^{-x(s-1)} < e^0 = 1.$$

The conditions in the bounded convergence theorem (2.265) are fulfilled, and termwise integration can be performed. Recalling the gamma function (1.38), using $\Gamma(n) = (n-1)!$ *for* $n \in \mathbb{N}$, *and the easy-to-verify (use* $u = mx$*)*

$$I = \int_0^\infty x^n \mathrm{e}^{-mx} dx = m^{-1} \int_0^\infty (u/m)^n \mathrm{e}^{-u} du = m^{-(n+1)} \Gamma(n+1), \quad m > 0,$$

and (2.264), this gives

$$\begin{aligned} \int_0^\infty e^{-sx} \frac{\sin x}{x} dx &= \sum_{n=0}^\infty \frac{(-1)^n}{(2n+1)!} \int_0^\infty e^{-sx} x^{2n} dx = \sum_{n=0}^\infty \frac{(-1)^n}{(2n+1)!} \frac{\Gamma(2n+1)}{s^{2n+1}} \\ &= \sum_{n=0}^\infty \frac{(-1)^n}{2n+1} \left(\frac{1}{s}\right)^{2n+1} = \arctan\left(s^{-1}\right), \end{aligned} \tag{2.271}$$

which is required in Example 2.54. ■

We now turn to the conditions that allow for interchange of limits and differentiation, beginning with some illustrations of what conditions are *not* sufficient.

Example 2.90 *Let* $f_n(x) = (\sin nx)/n$ *so that,* $\forall x \in \mathbb{R}$, $\lim_{n\to\infty} f_n(x) = 0 =: f(x)$. *Then* $f'(x) = 0$ *but* $f_n'(x) = \cos nx$ *and,* $\forall n \in \mathbb{N}$, $f_n'(0) = 1$, *so that* $\exists x \in \mathbb{R}$ *such that* $\lim_{n\to\infty} \frac{d}{dx} f_n(x) \neq \frac{d}{dx} \lim_{n\to\infty} f_n(x)$. *Given the previous results on interchange of limit and integral, one might expect that uniform convergence is sufficient. But,* $\forall x \in \mathbb{R}$,

$$|f_n(x) - f_m(x)| = \left|\frac{\sin nx}{n} - \frac{\sin mx}{m}\right| \leq \left|\frac{1}{n} - \frac{-1}{m}\right| = \left|\frac{1}{n} + \frac{1}{m}\right|,$$

so $\forall \epsilon > 0$, $\exists N \in \mathbb{N}$ *such that,* $\forall n, m > N$, $|f_n(x) - f_m(x)| < \epsilon$, *i.e.,* f_n *is uniformly Cauchy and, by (2.244),* f_n *is uniformly convergent. Thus, uniform convergence is not enough to ensure the interchange of limit and derivative.*

Observe that $f_n'(x) = \cos nx$ *is not convergent (pointwise or uniformly). It turns out that uniform convergence of* $\{f_n'\}$ *is necessary for interchange.* ■

Example 2.91 *Let* $I = [-1, 1]$ *and* $f_n(x) = (x^2 + n^{-1})^{1/2}$ *for* $x \in I$, *so that* $f_n \in \mathcal{C}^1$ *with* $f_n'(x) = x(x^2 + n^{-1})^{-1/2}$. *Figure 12 shows* f_n *and* f_n' *for several* n. *In the limit,* $f_n(x) \to f(x) := |x|$, *which is not differentiable at* $x = 0$. *In fact,* $f_n(x) \rightrightarrows f(x)$, *because, as shown next,* $m_n = \max_{x\in I} |f_n(x) - f(x)| = n^{-1/2}$ *and result (2.246).*

To derive m_n, *first note that* $f_n(x) - f(x)$ *is symmetric in* x, *so we can restrict attention to* $x \in [0, 1]$, *in which case* $d(x) = |f_n(x) - f(x)| = (x^2 + n^{-1})^{1/2} - x > 0$, *for* $x \in [0, 1]$. *Its first derivative is* $d'(x) = x(x^2 + n^{-1})^{-1/2} - 1$, *which is strictly negative for all* $x \in [0, 1]$ *and* $n \in \mathbb{N}$. *(At* $x = 1$, $d'(x) = \sqrt{n/(n+1)} - 1$.*) Thus,* $d(x)$ *reaches its maximum on* $[0, 1]$ *at* $x = 0$, *so that*

$$\max_{x\in[0,1]} |d(x)| = \max_{x\in I} |f_n(x) - f(x)| = d(0) = n^{-1/2}.$$

Also, $f_n'(x) \to x/|x|$ *for* $x \neq 0$, *but the convergence cannot be uniform at* $x = 0$, *because, for any* $n \in \mathbb{N}$,

$$\begin{aligned} \lim_{x\to 0^+} \left|f_n'(x) - \frac{x}{|x|}\right| &= \lim_{x\to 0^+} \left|\frac{x}{\sqrt{x^2 + n^{-1}}} - \frac{x}{|x|}\right| = \lim_{x\to 0^+} \frac{x}{|x|} - \lim_{x\to 0^+} \frac{x}{\sqrt{x^2 + n^{-1}}} \\ &= 1 - \lim_{x\to 0^+} \frac{x/x}{\sqrt{x^2/x^2 + 1/nx^2}} = 1 - 0 = 1. \end{aligned}$$ ■

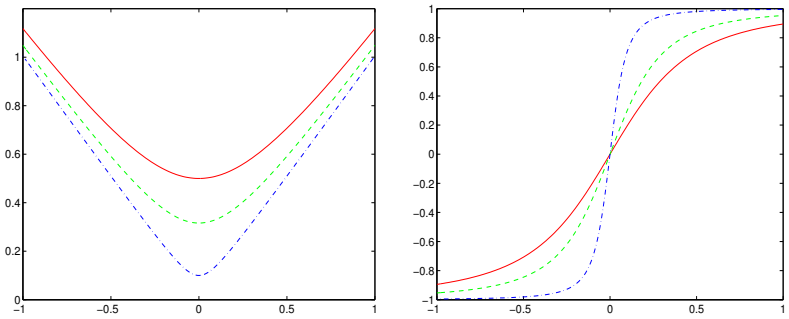


Figure 12: Function $f_n(x) = (x^2 + n^{-1})^{1/2}$ (left) and f_n' (right), for $n = 4$ (solid), $n = 10$ (dashed) and $n = 100$ (dash-dot).

Let $f : D \to \mathbb{R}$, where $I = [a, b] \subset D$. The following theorem gives the desired result, recalling from §2.2.1 that $\mathcal{C}^1$ is the class of continuously differentiable functions, i.e., f is differentiable and $f'(x)$ is continuous on D.

Let $f_n \in \mathcal{C}^1(I)$ such that $f_n'(x) \rightrightarrows g(x)$ and $f_n(x) \to f(x)$ on I. Then

$$g \in \mathcal{C}^0(I) \text{ and, } \forall x \in I,\ f'(x) = g(x). \tag{2.272}$$

Proof: That $g \in \mathcal{C}^0$ follows directly from (2.245). For $a, x \in I$, (2.130) and FTC (2.143) imply that, $\forall n \in \mathbb{N}$, $\int_a^x f_n' = f_n(x) - f_n(a)$. As $f_n'(x) \rightrightarrows g(x)$, taking the limit as $n \to \infty$ and using (2.254) gives $\int_a^x g = f(x) - f(a)$. As $g \in \mathcal{C}^0$, differentiating this via FTC (2.146) yields $g(x) = f'(x)$.

Example 2.92 *Again consider the geometric series $S(x) = \sum_{k=1}^{\infty} x^k$, which, from (2.214), converges pointwise for $x \in (-1, 1)$ to $S(x) = x/(1-x)$. That is, with $S_n(x) = \sum_{k=1}^{n} x^k$ the nth partial sum, $S_n(x) \to S(x)$, and, being a polynomial, $S_n \in \mathcal{C}^1(I)$ for $I = [-1, 1]$. To apply (2.272), we need to show that $\exists g$ such that $S_n'(x) \rightrightarrows g(x)$. With $S_n'(x) = \sum_{k=1}^{n} k x^{k-1}$, for $r \in (0, 1)$ and $\epsilon > 0$ such that $r + \epsilon < 1$, the binomial theorem (1.21) implies*

$$(r+\epsilon)^k = \sum_{i=0}^{k} \binom{k}{i} \epsilon^i r^{k-i} = r^k + k\epsilon r^{k-1} + \cdots + \epsilon^k. \tag{2.273}$$

The positivity of the terms on the rhs of (2.273) imply that $k\epsilon r^{k-1} < (r+\epsilon)^k$, so that the Weierstrass M-test implies that $\epsilon \sum_{k=1}^{\infty} k r^{k-1}$ is also uniformly convergent for $x \in [-r, r]$.

Thus, from (2.272) with $g(x) = S'(x) = \frac{d}{dx}\left(\lim_{n\to\infty} \sum_{k=1}^{n} x^k\right)$,

$$S_n'(x) = \sum_{k=1}^{n} k x^{k-1} \rightrightarrows g(x) = \frac{d}{dx}\left(\frac{x}{1-x}\right) = \frac{1}{(1-x)^2}. \tag{2.274}$$

As this holds $\forall x \in [-r, r]$, where r is an arbitrary number from the open interval $(0, 1)$, (2.274) holds for all $|x| < 1$. (If this were not true, then there would exist an $x \in (0, 1)$, say x_0, for which it were not true, but the previous analysis applies to all $x \in [0, x_0 + \epsilon]$, where ϵ is such that $x_0 + \epsilon < 1$, which always exists.) Thus, $S'(x) = (1-x)^{-2}$ on $(-1, 1)$.

To add some intuition and informality, let

$$D = \lim_{n\to\infty} \sum_{k=1}^{n} k x^{k-1} = 1 + 2x + 3x^2 + \cdots \quad \text{and} \quad xD = x + 2x^2 + 3x^3 + \cdots,$$

so that $D - xD = 1 + x + x^2 + \cdots = \sum_{k=0}^{\infty} x^k = (1-x)^{-1}$*, which, for* $|x| < 1$*, converges, and* $D = (1-x)^{-2}$*. Also see the next section.* ■

The assumptions in result (2.272) can be somewhat relaxed, though we won't require it. In particular, as proven, e.g., in Stoll (2001, pp. 340-1) and other analysis books:

> **Let** $\{f_n\}$ **be a sequence of differentiable functions on** $I = [a, b]$**. If** $f_n'(x) \rightrightarrows g(x)$ **on** I **and** $\exists\, x_0 \in I$ **such that** $\{f_n(x_0)\}$ **converges, then** $f_n(x) \rightrightarrows f(x)$ **on** I**, and,** $\forall x \in I$, $f'(x) = g(x)$.

2.5.6 Power and Taylor Series

> I regard as quite useless the reading of large treatises of pure analysis: too large a number of methods pass at once before the eyes. It is in the works of applications that one must study them. (Joseph-Louis Lagrange)

A series of the form $\sum_{k=0}^{\infty} a_k x^k$ for sequence $\{a_k\}$ is a *power series* in x with coefficients a_k. More generally, $S(x) = \sum_{k=0}^{\infty} a_k (x-c)^k$ is a power series in $(x-c)$, where $c \in \mathbb{R}$. With $f_k = a_k (x-c)^k$, the exponential growth rate of S is $g(x) = \limsup |f_k|^{1/k} = |x-c| \limsup |a_k|^{1/k}$, and, from the root test, S converges absolutely if $g(x) < 1$ and diverges for $g(x) > 1$. We define:

$$\textbf{The } \textit{\textbf{radius of convergence}} \textbf{ of } S \textbf{ is } R = 1/\limsup |a_k|^{1/k}. \tag{2.275}$$

As $1/0 = \infty$, $R = \infty$ if $\limsup |a_k|^{1/k} = 0$, and, similarly, as $1/\infty = 0$, $R = 0$ if $\limsup |a_k|^{1/k} = \infty$.

Power series $S(x)$ converges if $g(x) = |x-c| R^{-1} < 1$, or

$$S(x) \textbf{ converges if } |x-c| < R\textbf{, and diverges if } |x-c| > R. \tag{2.276}$$

When working with series involving factorials, the following result becomes very useful: If $\lim_{k\to\infty} |a_{k+1}/a_k|$ exists, then, from (2.205), $\liminf |a_{k+1}/a_k| = \limsup |a_{k+1}/a_k|$ and, from (2.234), these equal $\limsup |a_k|^{1/k}$, so that the radius of convergence is $R = 1/\lim_{k\to\infty} |a_{k+1}/a_k|$.

Example 2.93 *Consider the power series of the form*

$$S(x) = \sum_{k=0}^{\infty} a_k x^k, \quad \textit{where} \quad a_k = \frac{(-1)^k}{m^k (k!)^p} \quad \textit{and} \quad m, p \in \mathbb{R}_{>0}.$$

As

$$\lim_{k\to\infty} \frac{|a_{k+1}|}{|a_k|} = \lim_{k\to\infty} \frac{m^k (k!)^p}{m^{k+1} ((k+1)!)^p} = \lim_{k\to\infty} \frac{1}{m (1+k)^p} = 0,$$

we have $R = \infty$. ■

The next result relates power series and uniform convergence.

> **If power series** S **has radius of convergence** $R > 0$**, then,** $\forall b \in (0, R)$,
>
> $$S \textbf{ converges uniformly for all } x \textbf{ with } |x-c| \le b. \tag{2.277}$$

Proof: Choose $\epsilon > 0$ such that $b + \epsilon \in (b, R)$, which implies $\limsup |a_k|^{1/k} = R^{-1} < (b+\epsilon)^{-1}$. From (2.203), $\exists N \in \mathbb{N}$ such that, $\forall n \geq N$, $|a_k|^{1/k} < (b+\epsilon)^{-1}$, so that, $\forall n \geq N$ and $|x-c| \leq b$, $\left|a_k (x-c)^k\right| \leq |a_k| b^k < (b/(b+\epsilon))^k$. As $\sum_{k=1}^{\infty} (b/(b+\epsilon))^k < \infty$, the result follows from the Weierstrass M-test.

Example 2.94 *In Example 2.92, the uniform convergence of $\sum_{k=1}^{\infty} kx^{k-1}$ was shown via the binomial theorem and the Weierstrass M-test. The following way is easier: As $\sum_{k=1}^{\infty} kx^{k-1} = \sum_{j=0}^{\infty} (j+1) x^j$, let $a_j = j+1$, so that, from a small extension of the first limit result in Example 2.10, $\limsup |a_j|^{1/j} = \lim_{j\to\infty} (j+1)^{1/j} = 1$, and $R = 1$. Thus, $\sum_{k=1}^{\infty} kx^{k-1}$ converges, from (2.276) with $c = 0$, for $x \in (-1, 1)$, and (2.277) implies that $\sum_{k=1}^{\infty} kx^{k-1}$ is uniformly convergent on $[-r, r]$ for each $r \in (0, 1)$.* ■

Theorem (Abel): **Suppose $S(x) = \sum_{k=0}^{\infty} a_k x^k$ has radius of convergence $R = 1$. If $\sum_{k=0}^{\infty} a_k < \infty$, then**

$$\lim_{x\to 1^-} S(x) = S(1) = \sum_{k=0}^{\infty} a_k.$$

See, e.g., Goldberg (1964, §9.6) or Stoll (2001, §8.7) for proof. Naturally, Abel's theorem can also be stated for general c and $R > 0$.

Example 2.95 *Let $S(x) = \sum_{k=1}^{\infty} (-1)^{k+1} x^k / k$. From Example 2.9, $\lim |1/k|^{1/k} = 1$. Results (2.205) and (2.275) then imply that the radius of convergence of S is $R = 1$. From the alternating series test (*Dirichlet Test*, page 105), $S(1) = \sum_{k=1}^{\infty} (-1)^{k+1}/k$ is also convergent. Abel's theorem and (2.261) thus imply that*

$$\ln 2 = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots. \tag{2.278}$$

Another interesting method of proof of (2.278) is given in Loya (2017, §4.7.5), along with other expressions for it. ■

Example 2.96 *From Example 2.86,*

$$S(y) = \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} y^{2n+1} = \arctan(t), \quad t \in [0, 1).$$

From the alternating series test, $S(1)$ converges, so that, from Abel's theorem,

$$\frac{\pi}{4} = \arctan(1) = \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots$$

giving a (rather inefficient) method of calculating pi. ■

Of great use is the following result for termwise differentiation of power series.

Let $f(x) = \sum_{k=0}^{\infty} a_k (x-c)^k$ for $|x-c| < R$, where $R > 0$ is the radius of convergence of f. Then $d(x) = \sum_{k=1}^{\infty} k a_k (x-c)^{k-1}$ has radius of convergence R and

$$f'(x) = d(x) \text{ for } x \text{ such that } |x-c| < R. \tag{2.279}$$

Proof: Using the two limit results in Example 2.10, the exponential growth rate of $d(x)$ is, for $x \neq c$,

$$\begin{aligned}
\limsup \left| k a_k (x-c)^{k-1} \right|^{1/k} &= \limsup |k|^{1/k} \limsup \left| (x-c)^{k-1} \right|^{1/k} \limsup |a_k|^{1/k} \\
&= 1 \cdot \limsup \left| \frac{(x-c)^k}{x-c} \right|^{1/k} \limsup |a_k|^{1/k} \\
&= \frac{|x-c|}{\lim_{k\to\infty} |x-c|^{1/k}} \limsup |a_k|^{1/k} = |x-c| \limsup |a_k|^{1/k},
\end{aligned}$$

so that $d(x)$ has the same radius of convergence as does $f(r)$, namely R. That $f'(x) = d(x)$ for $|x-c| < R$ follows directly from the results in (2.277) and (2.272).

Thus, the result in Example 2.92 could have been obtained immediately via (2.279). It also implies that, if $f(x) = \sum_{k=0}^{\infty} a_k (x-c)^k$ and $g(x) = \sum_{k=0}^{\infty} b_k (x-c)^k$ are power series with radius of convergence R that are equal for $|x-c| < R$, then $a_k = b_k$ for $k = 0, 1, \ldots$. To see this, note that $f, g \in \mathcal{C}^\infty$ so $f^{(n)}(x) = g^{(n)}(x)$ for $n \in \mathbb{N}$ and $|x-c| < R$. In particular,

$$f^{(n)}(x) = \sum_{k=n}^{\infty} k(k-1)\cdots(k-n+1) a_k (x-c)^{k-n}$$

and, as $0^0 = 1$, $f^{(n)}(c) = n! a_n$. Thus, $n! a_n = f^{(n)}(c) = g^{(n)}(c) = n! b_n$ for $n = 0, 1, \ldots$, which implies that $a_n = b_n$ for $n = 0, 1, \ldots$.

Repeated use of (2.279) implies that $f \in \mathcal{C}^\infty(x-R, x+R)$, i.e., that f is infinitely differentiable on $(x-R, x+R)$. The converse, however, does not hold, i.e., there exist functions in $\mathcal{C}^\infty(I)$ that cannot be expressed as a power series for particular $c \in I$. The ubiquitous example is to use $c = 0$ and the function given by $f(x) = \exp(-1/x^2)$ for $x \neq 0$ and $f(0) = 0$. Let I be an open interval. A function $f : I \to \mathbb{R}$ is said to be *analytic* in I if, $\forall c \in I$, there exists a sequence $\{a_k\}$ in $\mathbb{R}$ and a $\delta > 0$ such that, $\forall x$ with $|x-c| < \delta$, $f(x) = \sum_{k=0}^{\infty} a_k (x-c)^k$. Thus, the class of analytic functions is a proper subset of $\mathcal{C}^\infty$.

Recall from (2.33) that, for a differentiable function f, $f(x+h) \approx f(x) + h f'(x)$, accurate for h near zero, i.e., knowledge of a function and its derivative at a specfied point, x, can be used to approximate the function at other points near x. By replacing x with c and then setting $h = x - c$, this can be written as

$$f(x) \approx f(c) + (x-c) f'(c). \tag{2.280}$$

For example, with $f(x) = e^x$ and $c = 0$, (2.280) reads $e^x \approx e^0 + x e^0 = 1 + x$, which is accurate for $x \approx 0$. When evaluated at $x = c$, (2.280) is exact, and taking first derivatives of both sides w.r.t. x gives $f'(x) \approx f'(c)$, which is again exact at $x = c$. One might imagine that accuracy is improved if terms involving higher derivatives are taken into account. This is the nature of a *Taylor polynomial*, which was developed by Brooks Taylor, 1685–1731 (though variants were independently discovered by others, such as Gregory, Newton, Leibniz, Johann Bernoulli and de Moivre). It was only in 1772 that Joseph-Louis Lagrange (1736–1813) recognized the importance of the contribution, proclaiming it the basic principle of the differential calculus. Lagrange is also responsible for characterizing the error term. The first usage of the term Taylor series appears to be by Simon Lhuilier (1750–1840) in 1786.

Let $f : I \to \mathbb{R}$, where I is an open interval, and let $c \in I$. If $f^{(n)}(x)$ exists for all $x \in I$, then the nth order *Taylor polynomial* of f at c is

$$T_n(x; f, c) = T_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(c)}{k!} (x-c)^k, \tag{2.281}$$

and if $f \in \mathcal{C}^\infty(I)$, then the *Taylor series* of f at c is

$$T(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(c)}{k!} (x-c)^k. \tag{2.282}$$

When $c = 0$, (2.282) is also referred to as the *Maclaurin series*, after Colin Maclaurin (1698–1746). As in (2.280), observe that $T_n(c) = f(c)$, $T_n'(c) = T_n'(x)|_{x=c} = f'(c)$, up to $T_n^{(r)}(c) = f^{(r)}(c)$, so that locally (i.e., for x near c), $T_n(c)$ behaves similarly to $f(x)$ and could be used for effective approximation. The *remainder* between f and $T_n(x)$ is defined as $R_n(x) := f(x) - T_n(x)$, and *Taylor's formula with remainder* is given by

$$f(x) = T_n(x) + R_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(c)}{k!} (x-c)^k + R_n(x). \tag{2.283}$$

Clearly, $f(x) = T(x)$ iff $\lim_{n\to\infty} R_n(x) = 0$. If $f^{(n+1)}(x)$ exists for all $x \in I$, then the *Lagrange form of the remainder* is

$$R_n(x) = \frac{f^{(n+1)}(\zeta)}{(n+1)!} (x-c)^{n+1}, \quad \zeta \text{ between } x \text{ and } c, \tag{2.284}$$

and from (2.4), $\lim_{n\to\infty} R_n(x) = 0$.

Another useful way of expressing (2.283) and (2.284) is (obtained by replacing c with x, and x with $x+h$, where $x \in I$ and h is small enough such that the "perturbation" $x+h \in I$; and we also switch from n to k, because we will refer to this formula in the multivariate section, and n is reserved for something else)

$$f(x+h) - \left[f(x) + f'(x)h + \cdots + \left(\frac{1}{k!}\right) f^k(x) h^k \right] = \frac{f^{k+1}(x+\theta h)}{(k+1)!} \cdot h^{k+1} \tag{2.285}$$

for some $0 < \theta < 1$, this being the analog of ζ between x and c. Dividing by h^k shows that

$$\lim_{h\to 0} \frac{f(x+h) - \left[f(x) + f'(x)h + \cdots + (1/k!) f^k(x) h^k \right]}{h^k} = 0. \tag{2.286}$$

Observe that, for $n = 1$, we can express (2.280) as

$$f(x) = f(c) + f'(c)(x-c) + r(x), \quad r(x) = \frac{1}{2} f''(\zeta) (x-c)^2, \tag{2.287}$$

with (as $f''(\zeta)$ exists, by assumption)

$$\lim_{x\to c} \frac{r(x)}{|x-c|} = \frac{1}{2} f''(\zeta) \lim_{x\to c} \frac{(x-c)^2}{|x-c|} = \frac{1}{2} f''(\zeta) \lim_{x\to c} |x-c| = 0. \tag{2.288}$$

Thus, $f(c) + f'(c)(x-c)$ is an (affine, for $c \neq 0$) linear approximation to $f(x)$ with the property that, not only is the error term $r(x)$ in (2.287) such that $\lim_{x\to c} r(x) = 0$, but also

(2.288), i.e., the limit of $r(x)$ after dividing by the linear quantity that itself goes to zero, is zero. In this sense, it is the best linear approximation to the function f at the point c. Representation (2.287) and (2.288) will be of use in the multivariate function case, when, in two dimensions, $T_1(x)$ in (2.281) will have two terms, one for each of the two input variables, and will represent the "best" affine two-dimensional plane approximation to a differentiable (defined in (4.31)) function, at a point (x_0, y_0) in the interior of its domain.

To prove (2.284), as in Bartle and Sherbert (1st edition, 1983, p. 222),[23] assume $x \neq c$ and let $J = [x, c] \cup [c, x]$, i.e., $J = [x, c]$ if $x < c$, and $[c, x]$ if $c < x$. Then, for $t \in J$, let

$$P_n(t) := f(x) - f(t) - (x - t) f'(t) - \frac{(x-t)^2 f''(t)}{2!} \cdots - \frac{(x-t)^n f^{(n)}(t)}{n!}, \quad (2.289)$$

with $P_n(x) = 0$. Then $P_1'(t) = -f'(t) - [(x-t) f''(t) + f'(t)(-1)] = -(x-t) f''(t)$, which can be written as $-(x-t)^n f^{(n+1)}(t)/n!$ for $n = 1$. Now use induction: assume this holds for $n-1$; then

$$\begin{aligned} P_n'(t) &= \frac{d}{dt}\left(P_{n-1}(t) - \frac{(x-t)^n f^{(n)}(t)}{n!}\right) \\ &= -\frac{(x-t)^{n-1} f^{(n)}(t)}{(n-1)!} - \frac{(x-t)^n f^{(n+1)}(t) - f^{(n)}(t)\, n (x-t)^{n-1}(-1)}{n!} \\ &= -\frac{(x-t)^n f^{(n+1)}(t)}{n!}. \end{aligned}$$

Now let

$$G(t) := P_n(t) - \left(\frac{x-t}{x-c}\right)^{n+1} P_n(c), \quad t \in J,$$

so that $G(c) = 0$ and $G(x) = P_n(x) = 0$. The mean value theorem then implies that there exists a $\zeta \in J$ (actually, the interior of J) such that

$$\frac{G(c) - G(x)}{c - x} = G'(\zeta),$$

so that $0 = G'(\zeta) = P_n'(\zeta) + (n+1)\frac{(x-\zeta)^n}{(x-c)^{n+1}} P_n(c)$. Thus,

$$\begin{aligned} P_n(c) &= -\frac{1}{n+1}\frac{(x-c)^{n+1}}{(x-\zeta)^n} P_n'(\zeta) = \frac{1}{n+1}\frac{(x-c)^{n+1}}{(x-\zeta)^n}\frac{(x-\zeta)^n f^{(n+1)}(\zeta)}{n!} \\ &= \frac{f^{(n+1)}(\zeta)}{(n+1)!}(x-c)^{n+1}, \end{aligned}$$

and (2.289) reads

$$\frac{f^{(n+1)}(\zeta)}{(n+1)!}(x-c)^{n+1} = f(x) - f(c) - (x-c) f'(c) - \frac{(x-c)^2 f''(c)}{2!} \cdots - \frac{(x-c)^n f^{(n)}(c)}{n!},$$

as was to be shown.

This proof, like other variants of it, are somewhat "rabbit-out-of-the-hat", in the sense that it is not at all clear how one stumbles upon choosing $P_n(t)$ and $G(t)$. Such elegant proofs

are just the result of concerted effort and much trial and error, and abound in mathematics, old and new. Indeed, referring to Gauss' style of mathematical proof, Niels Abel said that "He is like the fox, who effaces his tracks in the sand with his tail". In defense of his style, Gauss exclaimed that "no self-respecting architect leaves the scaffolding in place after completing the building". As encouragement, Gauss also said "If others would but reflect on mathematical truths as deeply and continuously as I have, then they would also make my discoveries".

For fun, we look at the "evolution" of the above proof, taken from Bartle and Sherbert (4th edition, 2011, p. 189). It is the same proof, but they shorten it and put more work on the reader:

6.4.1 Taylor's Theorem Let $n \in \mathbb{N}$, let $I := [a, b]$, and let $f : I \to \mathbb{R}$ be such that f and its derivatives $f', f'', \ldots, f^{(n)}$ are continuous on I and that $f^{(n+1)}$ exists on (a, b). If $x_0 \in I$, then for any x in I there exists a point c between x and x_0 such that

$$\begin{aligned} f(x) =& f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 \\ &+ \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + \frac{f^{(n+1)}(c)}{(n+1)!}(x - x_0)^{n+1}. \end{aligned}$$

Proof: Let x_0 and x be given and let J denote the closed interval with endpoints x_0 and x. We define the function F on J by

$$F(t) := f(x) - f(t) - (x - t)f'(t) - \cdots - \frac{(x - t)^n}{n!}f^{(n)}(t),$$

for $t \in J$. Then an easy calculation shows that we have

$$F'(t) = -\frac{(x - t)^n}{n!}f^{(n+1)}(t).$$

If we define G on J by

$$G(t) := F(t) - \left(\frac{x - t}{x - x_0}\right)^{n+1} F(x_0)$$

for $t \in J$, then $G(x_0) = G(x) = 0$. An application of Rolle's Theorem 6.2.3 yields a point c between x and x_0 such that

$$0 = G'(c) = F'(c) + (n + 1)\frac{(x - c)^n}{(x - x_0)^{n+1}}F(x_0).$$

Hence, we obtain

$$\begin{aligned} F(x_0) &= -\frac{1}{n+1}\frac{(x - x_0)^{n+1}}{(x - c)^n}F'(c) \\ &= \frac{1}{n+1}\frac{(x - x_0)^{n+1}}{(x - c)^n}\frac{(x - c)^n}{n!}f^{(n+1)}(c) = \frac{f^{(n+1)}(c)}{(n+1)!}(x - x_0)^{n+1}, \end{aligned}$$

which implies the stated result. Q.E.D.

The remainder $R_n(x)$ can be expressed in integral form, provided $f^{(n+1)}(x)$ exists for each $x \in I$ and, $\forall a, b \in I$, $f^{(n+1)} \in \mathcal{R}[a,b]$. In particular, and as proven in most all books on real analysis,

$$R_n(x) = \frac{1}{n!} \int_c^x f^{(n+1)}(t)(x-t)^n \, dt, \quad x \in I.$$

Example 2.97 *Let $f(x) = \sin x$, so that, from the conditions in (2.57), $f'(x) = \cos x$ and $f''(x) = -\sin x$. Thus, $f^{(2n)}(x) = (-1)^n \sin x$ and $f^{(2n+1)}(x) = (-1)^n \cos x$, for $x \in \mathbb{R}$ and $n \in \mathbb{N} \cup 0$. As $\sin 0 = 0$ and $\cos 0 = 1$, the nth order Taylor polynomial for $c = 0$ is thus*

$$T_n(x) = 0 + x - 0 - \frac{1}{6}x^3 + 0 + \frac{1}{120}x^5 + \cdots = \sum_{k=0}^{n} \frac{(-1)^k}{(2k+1)!} x^{2k+1}.$$

As $|\sin x| \leq 1$ and $|\cos x| \leq 1$, the remainder satisfies $|R_n(x)| \leq |x|^{n+1} / (n+1)!$, which goes to zero as $n \to \infty$. Thus,

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

which is its definition; see also Example 2.79. ■

3 Some Relevant Linear Algebra

> Do not worry about your problems with mathematics, I assure you, mine are far greater. (Albert Einstein)

This section borrows heavily from five fantastic books, recent and old, namely Anthony and Harvey, Linear Algebra, 2012; Lang, Calculus of Several Variables, 3rd ed., 1987; Flanigan and Kazdan, Calculus Two: Linear and Nonlinear Functions, 2nd ed., 1990; Shifrin and Adams, Linear Algebra: A Geometric Approach, 2nd ed., 2011; and Olver and Shakiban, Applied Linear Algebra, 2nd ed., 2018. The reader will note some repetition and redundancy. Terseness does not help students; what is useful is seeing various good approaches and presentations, noting their commonalities, and their differences. Blue text are additional comments from me.

3.1 (Hyper-)planes, Vector-Parametric and Cartesian Equations

We use bold face to denote a point, or n-tuple, in $\mathbb{R}^n$, e.g., $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, and also for multivariate mappings, e.g., $\mathbf{f} : \mathbb{R} \to \mathbb{R}^m$, $m > 1$. We assume the reader has a basic familiarity with n-tuples, row and column vectors, and basic operations with them, e.g., addition, transpose (denoted $\mathbf{x}^{\mathrm{T}}$ or $\mathbf{x}'$) and the inner (dot) product. In particular, for the latter:

Definition: For all column vectors $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$, and for all $\alpha \in \mathbb{R}$, the inner product is

$$\langle \mathbf{x}, \mathbf{y} \rangle := \mathbf{x}'\mathbf{y} = \mathbf{x} \cdot \mathbf{y} = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n, \tag{3.1}$$

and satisfies the following properties:
(i) $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$,
(ii) $\alpha \langle \mathbf{x}, \mathbf{y} \rangle = \langle \alpha\mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, \alpha\mathbf{y} \rangle$,
(iii) $\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$,
(iv) $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$, and $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ if and only if $\mathbf{x} = \mathbf{0}$.

Example 3.1 *(Flanigan and Kazdan, p. 77)*

(1) To prove that $\langle Z, X \rangle = 0$ for all $X \in \mathbb{R}^n$ implies $Z = 0$, let $X = Z$. Then $\langle Z, Z \rangle = 0$. But by property (iv), $Z = 0$.

(2) Prove that if $\langle Z_1, X \rangle = \langle Z_2, X \rangle$ for all $X \in \mathbb{R}^n$, then $Z_1 = Z_2$.
Proof: We have $\langle Z_1, X \rangle - \langle Z_2, X \rangle = 0$. From properties (ii) and (iii), for all X,

$$0 = \langle Z_1, X \rangle - \langle Z_2, X \rangle = \langle Z_1, X \rangle + \langle -Z_2, X \rangle = \langle Z_1 - Z_2, X \rangle .$$

By the first exercise, $Z_1 - Z_2 = 0$, and, thus, $Z_1 = Z_2$. ■

Definition: The length, or norm, or magnitude, of vector $\mathbf{a}$ is

$$\|\mathbf{a}\| = \sqrt{\langle \mathbf{a}, \mathbf{a} \rangle} = \langle \mathbf{a}, \mathbf{a} \rangle^{1/2} \geq 0. \tag{3.2}$$

Theorem (Polarization identity): For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\langle \mathbf{x}, \mathbf{y} \rangle = \frac{1}{4} \left(\|\mathbf{x} + \mathbf{y}\|^2 - \|\mathbf{x} - \mathbf{y}\|^2 \right), \tag{3.3}$$

expressing the standard inner product in terms of the norm.

Proof: Note that

$$\|\mathbf{x} \pm \mathbf{y}\|^2 = \langle \mathbf{x} \pm \mathbf{y}, \mathbf{x} \pm \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{x} \rangle \pm \langle \mathbf{x}, \mathbf{y} \rangle \pm \langle \mathbf{y}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 \pm 2\langle \mathbf{x}, \mathbf{y} \rangle.$$

Let $A = (a_1, a_2)$ be a point in the plane $\mathbb{R}^2$. We associate this point with the vector $\mathbf{a} = (a_1, a_2)^{\mathrm{T}}$, as representing a *displacement from the origin*, $(0, 0)$, to the point A. In this context, $\mathbf{a}$ is the position vector of the point A. Graphically, this displacement is illustrated by an arrow, or directed line segment, with the initial point at the origin and the terminal point at A. Even if a displacement does not begin at the origin, two displacements of the same length and the same direction are considered to be equal.

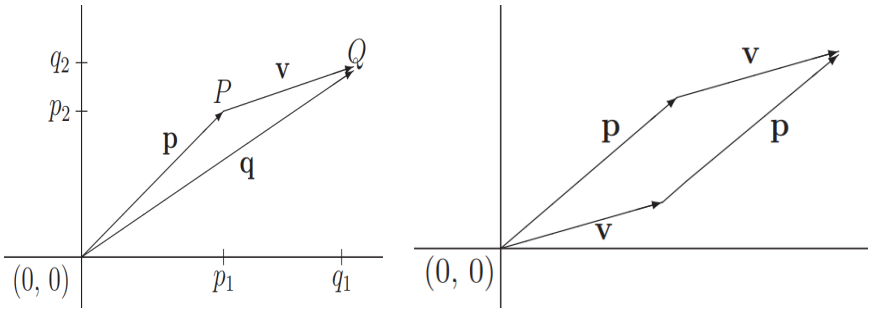


Figure 13: Left: Addition of two vectors. Right: The parallelogram law

If an object is displaced from the origin to a point P by the displacement $\mathbf{p}$, and then displaced from P to Q by the displacement $\mathbf{v}$, then the total displacement is given by the vector from 0 to Q, which is the position vector $\mathbf{q}$. So we would expect vectors to satisfy $\mathbf{q} = \mathbf{p} + \mathbf{v}$, both geometrically (in the sense of a displacement) and algebraically (by the definition of vector addition). This is true generally, for vectors in $\mathbb{R}^n$, and shown in the left panel of Figure 13. The order of displacements does not matter (similar to how the order of vector addition does not matter), so $\mathbf{q} = \mathbf{v} + \mathbf{p}$. For this reason, the addition of vectors is said to follow the parallelogram law; see the right panel of Figure 13. From the equation $\mathbf{q} = \mathbf{p} + \mathbf{v}$, we have $\mathbf{v} = \mathbf{q} - \mathbf{p}$. This is the displacement from P to Q. To help determine in which direction the vector $\mathbf{v}$ points, think of $\mathbf{v} = \mathbf{q} - \mathbf{p}$ as the vector that is added to the vector $\mathbf{p}$ in order to obtain the vector $\mathbf{q}$. The distance between points P and Q is (defined to be) $\|Q - P\| = \sqrt{(Q - P) \cdot (Q - P)} = \|\mathbf{v}\| = \langle \mathbf{v}, \mathbf{v} \rangle^{1/2}$.

Theorem (The Law of Cosines): For a triangle with sides a, b, c and opposite angles A, B, C, respectively, as pictured in Figure 14,

$$\begin{aligned} c^2 &= a^2 + b^2 - 2ab\cos C, \\ b^2 &= a^2 + c^2 - 2ac\cos B, \\ a^2 &= b^2 + c^2 - 2bc\cos A. \end{aligned} \tag{3.4}$$

Proof: To prove (3.4), use of Pythagoras and a basic trigonometric identity gives

$$\begin{aligned} c^2 &= (b - a\cos C)^2 + (0 - a\sin C)^2 \\ &= b^2 - 2ab\cos C + a^2\cos^2 C + a^2\sin^2 C \\ &= b^2 - 2ab\cos C + a^2\left(\cos^2 C + \sin^2 C\right) = a^2 + b^2 - 2ab\cos C. \end{aligned}$$

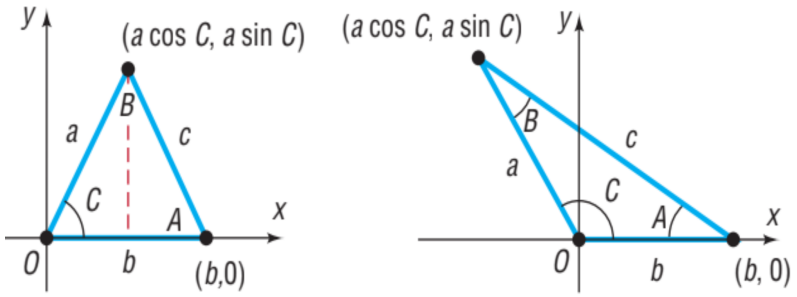


Figure 14: Triangles with C acute and obtuse, respectively. Taken from Sullivan, Trigonometry, 9th ed., 2012, p. 275

This can be used to provide the (very common) proof of the following crucial result:

Theorem: Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ and let θ denote the angle between them. Then

$$\langle \mathbf{a}, \mathbf{b} \rangle = \|\mathbf{a}\| \|\mathbf{b}\| \cos\theta. \tag{3.5}$$

Proof: As in Anthony and Harvey, the law of cosines states that $c^2 = a^2 + b^2 - 2ab\cos\theta$, where $c = \|\mathbf{b} - \mathbf{a}\|$, $a = \|\mathbf{a}\|$, $b = \|\mathbf{b}\|$. That is, $\|\mathbf{b} - \mathbf{a}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - 2\|\mathbf{a}\|\|\mathbf{b}\|\cos\theta$. Expanding the inner product and using its properties, we have

$$\|\mathbf{b} - \mathbf{a}\|^2 = \langle \mathbf{b} - \mathbf{a}, \mathbf{b} - \mathbf{a} \rangle = \langle \mathbf{b}, \mathbf{b} \rangle + \langle \mathbf{a}, \mathbf{a} \rangle - 2\langle \mathbf{a}, \mathbf{b} \rangle,$$

so that $\|\mathbf{b} - \mathbf{a}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - 2\langle \mathbf{a}, \mathbf{b} \rangle$. Comparing the two expressions yields (3.5).

We now show a different proof of (3.5), from Flanigan and Kazdan, using Figure 15.

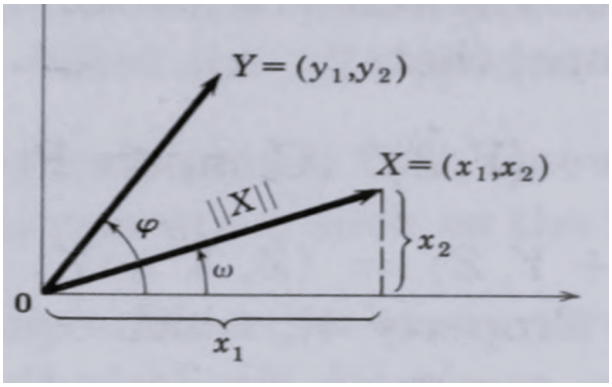


Figure 15: For the 2nd proof of (3.5). From Flanigan and Kazdan, p. 78.

Proof: Let θ be the angle between X and Y, i.e., $\theta = \varphi - \omega$. We know $\langle X, Y \rangle = x_1y_1 + x_2y_2$, where $X = (x_1, x_2)$ and $Y = (y_1, y_2)$ as usual. We will translate $x_1y_1 + x_2y_2$ into trigonometry. Let ω and φ be the angles from the horizontal axis to X and Y, respectively. Now we note that

$$\cos\omega = \frac{x_1}{\|X\|} \text{ and } \sin\omega = \frac{x_2}{\|X\|},$$

whence $x_1 = \|X\|\cos\omega$ and $x_2 = \|X\|\sin\omega$. Likewise $y_1 = \|Y\|\cos\varphi$ and $y_2 = \|Y\|\sin\varphi$. Thus

$$\langle X, Y \rangle = x_1y_1 + x_2y_2 = \|X\|\|Y\|(\cos\omega\cos\varphi + \sin\omega\sin\varphi).$$

Recall from (2.48) that $\cos\theta = \cos(\varphi - \omega) = \cos\omega\cos\varphi + \sin\omega\sin\varphi$. Thus $\langle X, Y\rangle = \|X\|\|Y\|\cos\theta$, as claimed.

Definition: The non-zero vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ are said to be *orthogonal*, or *perpendicular*, when the angle between them is $\theta = \pi/2$. As $\cos(\pi/2) = 0$, this is precisely when their inner product is zero. That is:

$$\textbf{The vectors a and b are orthogonal if and only if } \langle \mathbf{a}, \mathbf{b}\rangle = 0. \tag{3.6}$$

Definition: A *line in* $\mathbb{R}^n$ is given by a vector equation with one parameter of the form $\mathbf{x} = \mathbf{p} + t\mathbf{v}$, where $\mathbf{x}$ is the position vector of a point on the line, $\mathbf{p}$ is any particular point on the line, $\mathbf{v}$ is the direction of the line, and $t \in \mathbb{R}$.

If $\mathbf{p} = \mathbf{0}$, then line $\mathbf{x} = t\mathbf{v}$ goes through the origin, though note that a line with $\mathbf{p} \neq \mathbf{0}$ could still go through the origin. For $n = 3$,

$$\mathbf{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} + t \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}, \quad t \in \mathbb{R}. \tag{3.7}$$

In terms of Cartesian equations, equating components in (3.7) gives

$$x = p_1 + tv_1, \quad y = p_2 + tv_2, \quad z = p_3 + tv_3.$$

provided $v_i \neq 0$, $i = 1, 2, 3$, solving for t and equating,

$$\frac{x - p_1}{v_1} = \frac{y - p_2}{v_2} = \frac{z - p_3}{v_3}.$$

For example, to find Cartesian equations of the line

$$\mathbf{x} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + t \begin{pmatrix} -1 \\ 0 \\ 5 \end{pmatrix}, \quad t \in \mathbb{R},$$

equate components $x = 1 - t$, $y = 2$, $z = 3 + 5t$, and then solve for t in the first and third equation. The Cartesian equations are $1 - x = (z - 3)/5$ and $y = 2$, which is a line parallel to the xz-plane in $\mathbb{R}^3$.

Definition: A subset $\mathcal{S}$ of $\mathbb{R}^n$ is a *plane through the origin* (also called a two-dimensional linear subspace) if and only if it is the linear span of two vectors X, Y that do not lie on the same line through the origin; that is, if and only if

$$\mathcal{S} = \{Z \in \mathbb{R}^n : Z = \alpha X + \beta Y \text{ with } \alpha, \beta \in \mathbb{R}\}.$$

Definition: Vectors that lie on a common line through the origin are said to be collinear with the origin.

Thus the requirement on X and Y in the preceding definition is that X and Y be non-collinear with the origin. Two vectors in $\mathbb{R}^n$ that are non-collinear with the origin span a plane in $\mathbb{R}^n$. Further, a plane $\mathcal{S}$ is spanned by any two vectors that are in $\mathcal{S}$, and that are non-collinear with the origin.

If $\mathcal{S}$ is a subset of $\mathbb{R}^n$ and Z is a vector in $\mathbb{R}^n$, then we write

$$\mathcal{S} + Z = \{X' + Z : X' \in \mathcal{S}\}.$$

Definition: A subset $\mathcal{A}$ of $\mathbb{R}^n$ is called an *affine* subspace if and only if $\mathcal{A}$ is of the form $\mathcal{S} + Z$ for some linear subspace $\mathcal{S}$ and some vector Z in $\mathbb{R}^n$. In this case, $\mathcal{A}$ and $\mathcal{S}$ are said to be parallel. We call $\mathcal{A}$ a line if $\mathcal{S}$ is a line through the origin, or a plane if $\mathcal{S}$ is a plane through the origin.

Thus, dropping the requirement that a plane goes through the origin results in an affine subspace. The general definition (Flanigan and Kazdan, p. 53) is typical within the language of linear algebra:

Definition: A subset $\mathcal{A}$ of $\mathbb{R}^3$ is a plane if and only if $\mathcal{A}$ is the affine subspace consisting of the solutions of a linear equation $a_1x_1 + a_2x_2 + a_3x_3 = b$ with at least one of the coefficients a_1, a_2, and a_3 different from zero. Moreover, in this case, $\mathcal{A}$ is parallel to the two-dimensional linear subspace $\mathcal{S}$ of solutions of the homogeneous equation $a_1x_1 + a_2x_2 + a_3x_3 = 0$, that is, $\mathcal{A} = \mathcal{S} + Z$, where Z is any solution of $a_1x_1 + a_2x_2 + a_3x_3 = b$.

Definition: A plane in $\mathbb{R}^3$ is given by the *vector parametric equation*

$$\mathbf{x} = \mathbf{p} + s\mathbf{v} + t\mathbf{w}, \quad s, t, \in \mathbb{R}, \quad \mathbf{p}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^3,$$

provided that the vectors $\mathbf{v}$ and $\mathbf{w}$ are non-zero and are not parallel.

From (3.6), vector $\mathbf{x}$ is orthogonal to $\mathbf{n}$ if and only if $\langle \mathbf{n}, \mathbf{x} \rangle = 0$. This latter equation also characterizes the plane: If $\mathbf{n} = (a, b, c)^{\mathrm{T}}$ and $\mathbf{x} = (x, y, z)^{\mathrm{T}}$, then this equation can be written as

$$\langle \mathbf{n}, \mathbf{x} \rangle = \left\langle \begin{pmatrix} a \\ b \\ c \end{pmatrix}, \begin{pmatrix} x \\ y \\ z \end{pmatrix} \right\rangle = 0,$$

or

$$ax + by + cz = 0. \tag{3.8}$$

Definition: Form (3.8) is a *Cartesian equation* of a plane through the origin in $\mathbb{R}^3$.

Definition: The vector $\mathbf{n}$ is called a *normal vector to the plane.*

Any vector that is parallel to $\mathbf{n}$ will also be a normal vector and will lead to the same Cartesian equation. On the other hand, given any Cartesian equation of the form $ax + by + cz = 0$, this equation represents a plane through the origin in $\mathbb{R}^3$ with normal vector $\mathbf{n} = (a, b, c)^{\mathrm{T}}$.

To describe a plane that does not go through the origin, we choose a normal vector $\mathbf{n}$ and one point P on the plane with position vector $\mathbf{p}$. We then consider all displacement vectors that lie in the plane with initial point at P. If $\mathbf{x}$ is the position vector of any point on the plane, then the displacement vector $\mathbf{x} - \mathbf{p}$ lies in the plane, and $\mathbf{x} - \mathbf{p}$ is orthogonal to $\mathbf{n}$. Conversely, if the position vector $\mathbf{x}$ of a point satisfies $\langle \mathbf{n}, \mathbf{x} - \mathbf{p} \rangle = 0$, then the vector $\mathbf{x} - \mathbf{p}$ lies in the plane, so the point (with position vector $\mathbf{x}$) is on the plane. This is illustrated in Figure 16, albeit with different notation.

The orthogonality condition (3.6) means that the position vector of any point on the plane is given by the equation $\langle \mathbf{n}, \mathbf{x} - \mathbf{p} \rangle = 0$. Using properties of the inner product, we can rewrite this as $\langle \mathbf{n}, \mathbf{x} \rangle = \langle \mathbf{n}, \mathbf{p} \rangle$, where $\langle \mathbf{n}, \mathbf{p} \rangle = d$ is a constant. If $\mathbf{n} = (a, b, c)^{\mathrm{T}}$ and $\mathbf{x} = (x, y, z)^{\mathrm{T}}$, then $ax + by + cz = d$ is a Cartesian equation of a plane in $\mathbb{R}^3$. The plane goes through the

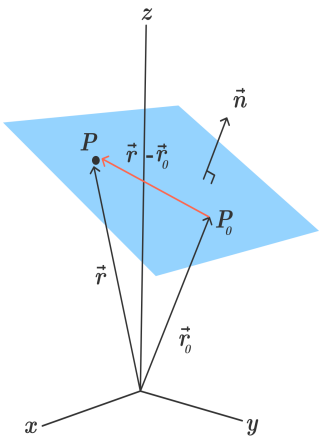


Figure 16: From https://brilliant.org/wiki/3d-coordinate-geometry-equation-of-a-plane/. Let P_0 be a point on the plane with position vector $\mathbf{r}_0$, and let P be some other point on the plane with position vector $\mathbf{r}$ (in place of $\mathbf{x}$). Then observe that the vector $\mathbf{r} - \mathbf{r}_0$ is the vector originating at P_0 and ending at P, and thus lies in the plane. Indeed, from the figure, note that $\mathbf{r} = \mathbf{r}_0 + (\mathbf{r} - \mathbf{r}_0)$. It is orthogonal to normal vector $\mathbf{n}$ as indicated.

origin if and only if $d = 0$. For example, the equation $2x - 3y - 5z = 2$ represents a plane that does not go through the origin, as $(x, y, z) = (0, 0, 0)$ does not satisfy the equation. To find a point on the plane, choose any two of the coordinates, say $y = 0$ and $z = 0$, implying $x = 1$, and that the point $(1, 0, 0)$ is on this plane. The components of a normal to the plane can be read from this equation as the coefficients of x, y, z : $\mathbf{n} = (2, -3, -5)^{\mathrm{T}}$.

The next example indicates one of a handful of "typical" questions involving this material, and most all books have something like it. Indeed, forthcoming Example 3.7 is similar. There, more explanation is provided as to why we need to solve the system of two equations (3.9); namely, for the plane $\mathcal{S}$, there must be a line orthogonal to $\mathcal{S}$, so we need a vector $\mathbf{a} = (a_1, a_2, a_3)$ that is orthogonal to both $\mathbf{u}$ and $\mathbf{v}$. It must satisfy (3.9).

Example 3.2 *(Flanigan and Kazdan, p. 53). Let us find an equation representing the plane $\mathcal{S} + Z$, where $Z = (3, 2, 1)$ and $\mathcal{S}$ is the plane through the origin that contains $\mathbf{u} = (0, 1, -3)$ and $\mathbf{v} = (1, 4, -1)$. We first find an equation representing $\mathcal{S}$ by solving the system*

$$a_2 - 3a_3 = 0, \qquad a_1 + 4a_2 - a_3 = 0, \tag{3.9}$$

for the unknowns a_1, a_2, a_3. Eliminate a_2 from the second equation to obtain the following solution set: $a_1 = -11a_3$, $a_2 = 3a_3$, and a_3 is arbitrary. We may choose $a_3 = 1$, which gives $a_1 = -11$ and $a_2 = 3$. Thus, $\mathcal{S}$ is the set of vectors (x_1, x_2, x_3) such that $-11x_1 + 3x_2 + x_3 = 0$. Now we want to find a inhomogeneous equation with the same coefficients such that Z is a solution. Plugging $Z = (z_1, z_2, z_3) = (3, 2, 1)$ into the left-hand side of the homogeneous equation, we get $(-11 \times 3) + (3 \times 2) + (1 \times 1) = -26$, so $\mathcal{S} + Z$ is the set of solutions of $-11x_1 + 3x_2 + x_3 = -26$. ■

The two representations, *vector parametric equation*, and *Cartesian equation*, are easily related, as shown in the next example.

Example 3.3 *(Anthony and Harvey, p. 42). Consider the plane*

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = s \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix} + t \begin{pmatrix} -2 \\ 1 \\ 7 \end{pmatrix} = s\mathbf{v} + t\mathbf{w}, \quad s, t \in \mathbb{R}. \tag{3.10}$$

To obtain a Cartesian equation in x, y and z, we equate the components in this vector equation,

$$x = s - 2t, \quad y = 2s + t, \quad z = -s + 7t,$$

and eliminate the parameters s and t. We begin by solving the first equation for s, and then substitute this into the second equation to solve for t in terms of x and y,

$$s = x + 2t \Rightarrow y = 2(x + 2t) + t = 2x + 5t \Rightarrow 5t = y - 2x \Rightarrow t = \frac{y - 2x}{5}.$$

We then substitute back into the first equation to obtain s in terms of x and y,

$$s = x + 2\left(\frac{y - 2x}{5}\right) \Rightarrow 5s = 5x + 2y - 4x \Rightarrow s = \frac{x + 2y}{5}.$$

Finally, we substitute for s and t in the third equation, $z = -s + 7t$, and simplify to obtain a Cartesian equation of the plane

$$3x - y + z = 0.$$

This Cartesian equation can be expressed as $\langle \mathbf{n}, \mathbf{x} \rangle = 0$, where

$$\mathbf{n} = \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

The vector $\mathbf{n}$ is a normal vector to the plane. We can check that $\mathbf{n}$ is, indeed, orthogonal to the plane by taking the inner product with the vectors $\mathbf{v}$ and $\mathbf{w}$, which lie in the plane.

Now consider displacing the plane so that it does not go through the origin, taking $\mathbf{p} = (3, 7, 2)'$, e.g.,

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 3 \\ 7 \\ 2 \end{pmatrix} + s \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix} + t \begin{pmatrix} -2 \\ 1 \\ 7 \end{pmatrix} = \mathbf{p} + s\mathbf{v} + t\mathbf{w}, \quad s, t \in \mathbb{R}, \tag{3.11}$$

which passes through the point $(3, 7, 2)$. Since the two planes (3.10) and (3.11) are parallel, they will have the same normal vectors, and the Cartesian equation of this plane is of the form $3x - y + z = d$. Since $(3, 7, 2)$ is a point on the plane, it must satisfy the equation for the plane. Substituting into the equation we find $d = 3(3) - (7) + (2) = 4$ (which is equivalent to finding d by using $d = \langle \mathbf{n}, \mathbf{p} \rangle$). So the Cartesian equation we obtain is $3x - y + z = 4$.

Conversely, starting with a Cartesian equation of a plane, we can obtain a vector equation. Consider the plane just discussed. We are looking for the position vector of a point on the plane whose components satisfy $3x - y + z = 4$, or, equivalently, $z = 4 - 3x + y$. (We can solve for any one of the variables x, y or z, but we chose z for simplicity.) So we are looking for all vectors $\mathbf{x}$ such that

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x \\ y \\ 4 - 3x + y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 4 \end{pmatrix} + x \begin{pmatrix} 1 \\ 0 \\ -3 \end{pmatrix} + y \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$

for any $x, y \in \mathbb{R}$. Therefore,

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 4 \end{pmatrix} + s \begin{pmatrix} 1 \\ 0 \\ -3 \end{pmatrix} + t \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \quad s, t \in \mathbb{R}$$

is a vector equation of the same plane as (3.11). It is difficult to spot at a glance that these two different vector equations in fact describe the same plane. The planes represented by the two vector equations have the same normal vector $\mathbf{n}$*, since the vectors* $(1,0,-3)^{\mathrm{T}}$ *and* $(0,1,1)^{\mathrm{T}}$ *are also orthogonal to* $\mathbf{n}$*. So we know that the two vector equations represent parallel planes. They are the same plane if they have a point in common. It is far easier to find values of* s *and* t *for which* $\mathbf{p} = (3,7,2)^{\mathrm{T}}$ *satisfies the new vector equation*

$$\begin{pmatrix} 3 \\ 7 \\ 2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 4 \end{pmatrix} + s \begin{pmatrix} 1 \\ 0 \\ -3 \end{pmatrix} + t \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \quad s,t \in \mathbb{R}$$

than the other way around (which is by showing that $(0,0,4)$ *satisfies the original equation) because of the positions of the zeros and ones in these direction vectors.* ■

Example 3.4 *(Anthony and Harvey, p. 46). The planes*

$$x + 2y - 3z = 0 \quad \textit{and} \quad x - 2y + 5z = 4$$

intersect in a line. The points of intersection are the points (x,y,z) *that satisfy both equations, so we solve the equations simultaneously. We begin by eliminating the variable* x *from the second equation, by subtracting the first equation from the second. This will naturally lead us to a vector equation of the line of intersection:*

$$\left.\begin{aligned} x + 2y - 3z &= 0 \\ x - 2y + 5z &= 4 \end{aligned}\right\} \Rightarrow \begin{aligned} x + 2y - 3z &= 0 \\ -4y + 8z &= 4. \end{aligned} \tag{3.12}$$

This last equation tells us that if $z = t$ *is any real number, then* $y = -1 + 2t$*. Substituting these expressions into the first equation, we find* $x = 2 - t$*. Then a vector equation of the line of intersection is*

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2 - t \\ -1 + 2t \\ t \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \\ 0 \end{pmatrix} + t \begin{pmatrix} -1 \\ 2 \\ 1 \end{pmatrix} =: \mathbf{p} + t\mathbf{v}.$$

This can be verified by showing that the point $\mathbf{p} = (2,-1,0)$ *satisfies both Cartesian equations, and that the vector* $\mathbf{v} = (-1,2,1)^{\mathrm{T}}$ *is orthogonal to the normal vectors of each of the planes (and therefore lies in both planes).* ■

In the previous example, we can envision the line induced by the two planes from Figure 17. The caption also indicates another way to compute a vector parallel to the intersecting line, namely by use of of the cross product of the two normal plane vectors. The cross product will be introduced below in §3.3, in (3.26). For two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^3$, it is given by

$$(x_2y_3 - x_3y_2,\ \ x_3y_1 - x_1y_3,\ \ x_1y_2 - x_2y_1)\,. \tag{3.13}$$

Based on the two normal vectors $(1,2,-3)$ and $(1,-2,5)$, this yields $(10-6,-3-5,-2-2) = (4,-8,-4)$, which is indeed parallel to $\mathbf{v} = (-1,2,1)^{\mathrm{T}}$. To recover point $\mathbf{p}$, note from the latter set of equations in (3.12) that x and y are the (in the common linear algebra language for Gaussian elimination) "basic" variables, while z is the "free" variable. Taking $z = 0$ (the simplest choice) implies $y = -1$ and $x = 2$, which is precisely point $\mathbf{p}$.

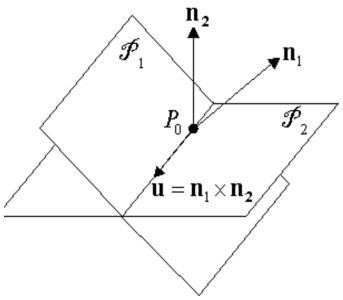


Figure 17: From https://math.stackexchange.com/questions/2387317. We have two non-parallel planes $\mathcal{P}_1$ and $\mathcal{P}_2$ with normal vectors $\mathbf{n}_1$ and $\mathbf{n}_2$, respectively. Let $\mathcal{L}$ be the line of intersection of $\mathcal{P}_1$ and $\mathcal{P}_2$. Let P_0 be a point on $\mathcal{L}$ and suppose that $\mathbf{v}$ is a vector parallel to $\mathcal{L}$. Note that $\mathbf{v}$ is a vector in both $\mathcal{P}_1$ and $\mathcal{P}_2$. This means that $\mathbf{v} \cdot \mathbf{n}_1 = 0$ and $\mathbf{v} \cdot \mathbf{n}_2 = 0$. That is, $\mathbf{v}$ is a vector orthogonal to both $\mathbf{n}_1$ and $\mathbf{n}_2$. Hence $\mathbf{v}$ is parallel to the cross-product $\mathbf{u} = \mathbf{n}_1 \times \mathbf{n}_2$.

Definition: The set of all points $(x_1, x_2, \ldots, x_n)$ that satisfy one Cartesian equation

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

is called a *hyperplane in* $\mathbb{R}^n$. That is, in $\mathbb{R}^n$, a hyperplane is an affine subspace of dimension $n - 1$.

In $\mathbb{R}^2$, a hyperplane is a line, and in $\mathbb{R}^3$ it is a plane, but for $n > 3$ we simply use the term hyperplane. The column vector $\mathbf{n} = (a_1, \ldots, a_n)'$ is a normal vector to the hyperplane. Writing the Cartesian equation in vector form, a hyperplane is the set of all vectors, $\mathbf{x} \in \mathbb{R}^n$ such that $\langle \mathbf{n}, \mathbf{x} - \mathbf{p} \rangle = 0$, where the normal vector $\mathbf{n}$ and the position vector $\mathbf{p}$ of a point on the hyperplane are given.

Example 3.5 *Using the Gauss-Jordan method, we find that the solution set of the system*

$$\begin{aligned} x_1 - x_2 + x_3 - 2x_4 &= -1 \\ x_2 + 3x_3 &= 0 \end{aligned}$$

is the set of vectors of the form

$$(x_1, x_2, x_3, x_4) = (-1, 0, 0, 0) + x_3(-4, -3, 1, 0) + x_4(2, 0, 0, 1),$$

with x_3 and x_4 arbitrary. (Here, x_1 and x_2 are the basic variables*, and x_3 and x_4 are the* free variables*.) Thus, the solution set is the two-dimensional affine subspace $\mathcal{A}$ of $\mathbb{R}^4$ that contains the vector $(-1, 0, 0, 0)$ and is parallel to the plane $\mathcal{S}$ through the origin spanned by the two linearly independent vectors $(-4, -3, 1, 0)$ and $(2, 0, 0, 1)$.* ■

3.2 Projection

We give presentations taken from three different textbooks.

3.2.1 Shifrin and Adams, Linear Algebra: A Geometric Approach

Starting with a two-dimensional picture of two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, where $\mathbf{y} \neq \mathbf{0}$, it suggests itself that we should be able to write $\mathbf{x}$ as the sum of a vector, $\mathbf{x}^{\parallel}$ (read "x-parallel"), that is a scalar multiple of $\mathbf{y}$ and a vector, $\mathbf{x}^{\perp}$ (read "x-perp"), that is orthogonal to $\mathbf{y}$. Let's suppose we have such an equation $\mathbf{x} = \mathbf{x}^{\parallel} + \mathbf{x}^{\perp}$, where $\mathbf{x}^{\parallel}$ is a scalar multiple of $\mathbf{y}$ and $\mathbf{x}^{\perp}$ is orthogonal to $\mathbf{y}$. To say that $\mathbf{x}^{\parallel}$ is a scalar multiple of $\mathbf{y}$ means that we can write $\mathbf{x}^{\parallel} = c\mathbf{y}$ for some scalar c. Now, assuming such an expression exists, we can determine c by taking the dot product of both sides of the equation with $\mathbf{y}$:

$$\mathbf{x} \cdot \mathbf{y} = \left(\mathbf{x}^{\parallel} + \mathbf{x}^{\perp}\right) \cdot \mathbf{y} = \left(\mathbf{x}^{\parallel} \cdot \mathbf{y}\right) + \left(\mathbf{x}^{\perp} \cdot \mathbf{y}\right) = \mathbf{x}^{\parallel} \cdot \mathbf{y} = (c\mathbf{y}) \cdot \mathbf{y} = c\|\mathbf{y}\|^2.$$

This means that

$$c = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{y}\|^2}, \quad \text{and so} \quad \mathbf{x}^{\parallel} = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{y}\|^2}\mathbf{y}.$$

The vector $\mathbf{x}^{\parallel}$ is called the projection of $\mathbf{x}$ onto $\mathbf{y}$, written $\text{proj}_{\mathbf{y}}\,\mathbf{x}$.

The fastidious reader may be puzzled by the logic here. We have apparently assumed that we can write $\mathbf{x} = \mathbf{x}^{\parallel} + \mathbf{x}^{\perp}$ in order to prove that we can do so. Of course, as it stands, this is not fair. Here's how we fix it. We now *define*

$$\mathbf{x}^{\parallel} = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{y}\|^2}\mathbf{y}, \quad \mathbf{x}^{\perp} = \mathbf{x} - \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{y}\|^2}\mathbf{y}.$$

Obviously, $\mathbf{x}^{\parallel} + \mathbf{x}^{\perp} = \mathbf{x}$ and $\mathbf{x}^{\parallel}$ is a scalar multiple of $\mathbf{y}$. All we need to check is that $\mathbf{x}^{\perp}$ is in fact orthogonal to $\mathbf{y}$. Well,

$$\begin{aligned}\mathbf{x}^{\perp} \cdot \mathbf{y} &= \left(\mathbf{x} - \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{y}\|^2}\mathbf{y}\right) \cdot \mathbf{y} = \mathbf{x} \cdot \mathbf{y} - \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{y}\|^2}\mathbf{y} \cdot \mathbf{y} \\ &= \mathbf{x} \cdot \mathbf{y} - \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{y}\|^2}\|\mathbf{y}\|^2 = \mathbf{x} \cdot \mathbf{y} - \mathbf{x} \cdot \mathbf{y} = 0,\end{aligned}$$

as required. Note that by finding a formula for c above, we have shown that $\mathbf{x}^{\parallel}$ is the unique multiple of $\mathbf{y}$ that satisfies the equation $\left(\mathbf{x} - \mathbf{x}^{\parallel}\right) \cdot \mathbf{y} = 0$.

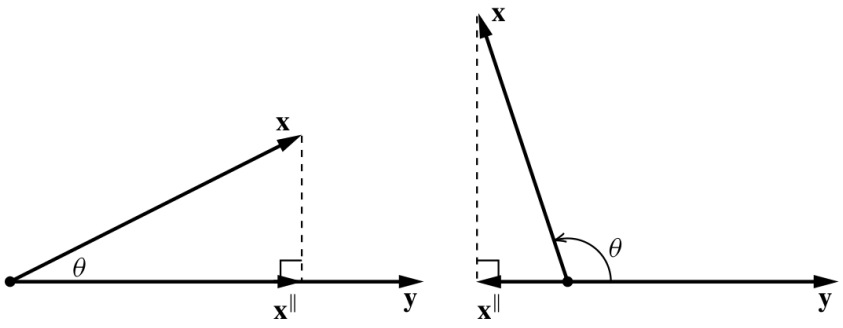


Figure 18: Projection and angle

Suppose $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$. We shall see next that the formula for the projection of $\mathbf{x}$ onto $\mathbf{y}$ enables us to calculate the angle between the vectors $\mathbf{x}$ and $\mathbf{y}$. Consider the right triangle

in Figure 18. Let θ denote the angle between the vectors $\mathbf{x}$ and $\mathbf{y}$. Remembering that the cosine of an angle is the ratio of the signed length of the adjacent side to the length of the hypotenuse, we see that

$$\cos\theta = \frac{\text{signed length of } \mathbf{x}^{\|}}{\text{length of } \mathbf{x}} = \frac{c\|\mathbf{y}\|}{\|\mathbf{x}\|} = \frac{\frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{y}\|^2}\|\mathbf{y}\|}{\|\mathbf{x}\|} = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}.$$

This, then, is the geometric interpretation of the dot product:

$$\mathbf{x}\cdot\mathbf{y} = \|\mathbf{x}\|\|\mathbf{y}\|\cos\theta. \tag{3.14}$$

Note that if the angle θ is obtuse, i.e., $\pi/2 < |\theta| < \pi$, then $c < 0$ (the signed length of $\mathbf{x}^{\|}$ is negative) and $\mathbf{x}\cdot\mathbf{y}$ is negative. Will this formula still make sense even when $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$? Geometrically, we simply restrict our attention to the plane spanned by $\mathbf{x}$ and $\mathbf{y}$ and measure the angle θ in that plane. This results in the following definition.

Definition: Let $\mathbf{x}$ and $\mathbf{y}$ be nonzero vectors in $\mathbb{R}^n$. We define the angle between them to be the unique θ satisfying $0 \le \theta \le \pi$ so that

$$\cos\theta = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}.$$

Example 3.6 *Consider the line ℓ_0 through the origin in $\mathbb{R}^2$ with direction vector $\mathbf{v} = (1, -3)$. The points on this line are all of the form*

$$\mathbf{x} = t(1, -3), \quad t \in \mathbb{R}.$$

Because $(3, 1)\cdot(1, -3) = 0$, we may take $\mathbf{a} = (3, 1)$ to be the normal vector to the line, and the Cartesian equation of ℓ_0 is

$$\mathbf{a}\cdot\mathbf{x} = 3x_1 + x_2 = 0.$$

As a check, suppose we start with $3x_1 + x_2 = 0$. Then we can write $x_1 = -\frac{1}{3}x_2$, and so the solutions consist of vectors of the form

$$\mathbf{x} = (x_1, x_2) = \left(-\frac{1}{3}x_2, x_2\right) = -\frac{1}{3}x_2(1, -3), \quad x_2 \in \mathbb{R}.$$

Letting $t = -\frac{1}{3}x_2$, we recover the original parametric equation.

Now consider the line ℓ passing through $\mathbf{x}_0 = (2, 1)$ with direction vector $\mathbf{v} = (1, -3)$. Then the points on ℓ are all of the form

$$\mathbf{x} = \mathbf{x}_0 + t\mathbf{v} = (2, 1) + t(1, -3), \quad t \in \mathbb{R}.$$

As promised, we take the same vector $\mathbf{a} = (3, 1)$ and compute that

$$3x_1 + x_2 = \mathbf{a}\cdot\mathbf{x} = \mathbf{a}\cdot(\mathbf{x}_0 + t\mathbf{v}) = \mathbf{a}\cdot\mathbf{x}_0 + t(\mathbf{a}\cdot\mathbf{v}) = \mathbf{a}\cdot\mathbf{x}_0 = (3, 1)\cdot(2, 1) = 7.$$

This is the Cartesian equation of ℓ. ■

The next example is similar to Example 3.2.

Example 3.7 *Consider the plane $\mathcal{P}_0$ passing through the origin spanned by $\mathbf{u} = (1, 0, 1)$ and $\mathbf{v} = (2, 1, 1)$. Our intuition suggests that there is a line orthogonal to $\mathcal{P}_0$, so we look for a vector $\mathbf{a} = (a_1, a_2, a_3)$ that is orthogonal to both $\mathbf{u}$ and $\mathbf{v}$. It must satisfy the equations*

$$\begin{aligned} a_1 + a_3 &= 0 \\ 2a_1 + a_2 + a_3 &= 0. \end{aligned}$$

Substituting $a_3 = -a_1$ into the second equation, we obtain $a_1 + a_2 = 0$, so $a_2 = -a_1$ as well. Thus, any candidate for $\mathbf{a}$ must be a scalar multiple of the vector $(1, -1, -1)$, and so we take $\mathbf{a} = (1, -1, -1)$ and try the equation

$$\mathbf{a} \cdot \mathbf{x} = (1, -1, -1) \cdot \mathbf{x} = x_1 - x_2 - x_3 = 0$$

for $\mathcal{P}_0$. Now, we know that $\mathbf{a} \cdot \mathbf{u} = \mathbf{a} \cdot \mathbf{v} = 0$. Does it follow that $\mathbf{a}$ is orthogonal to every linear combination of $\mathbf{u}$ and $\mathbf{v}$? We just compute: If $\mathbf{x} = s\mathbf{u} + t\mathbf{v}$, then

$$\begin{aligned} \mathbf{a} \cdot \mathbf{x} &= \mathbf{a} \cdot (s\mathbf{u} + t\mathbf{v}) \\ &= s(\mathbf{a} \cdot \mathbf{u}) + t(\mathbf{a} \cdot \mathbf{v}) = 0 \end{aligned}$$

as desired. As before, if we want the equation of the plane $\mathcal{P}$ parallel to $\mathcal{P}_0$ and passing through $\mathbf{x}_0 = (2, 3, -2)$, we take

$$\begin{aligned} x_1 - x_2 - x_3 = \mathbf{a} \cdot \mathbf{x} &= \mathbf{a} \cdot (\mathbf{x}_0 + s\mathbf{u} + t\mathbf{v}) \\ &= \mathbf{a} \cdot \mathbf{x}_0 + s(\mathbf{a} \cdot \mathbf{u}) + t(\mathbf{a} \cdot \mathbf{v}) \\ &= \mathbf{a} \cdot \mathbf{x}_0 = (1, -1, -1) \cdot (2, 3, -2) = 1. \end{aligned}$$

As this example suggests, a point $\mathbf{x}_0$ and a normal vector $\mathbf{a}$ give rise to the Cartesian equation of a plane in $\mathbb{R}^3$: $\mathbf{a} \cdot (\mathbf{x} - \mathbf{x}_0) = 0$, or, equivalently, $\mathbf{a} \cdot \mathbf{x} = \mathbf{a} \cdot \mathbf{x}_0$. Thus, every plane in $\mathbb{R}^3$ has an equation of the form $a_1x_1 + a_2x_2 + a_3x_3 = c$, where $\mathbf{a} = (a_1, a_2, a_3)$ is the normal vector and $c \in \mathbb{R}$.

Consider the set of points $\mathbf{x} = (x_1, x_2, x_3)$ defined by the equation $x_1 - 2x_2 + 5x_3 = 3$. Let's verify that this is, in fact, a plane in $\mathbb{R}^3$ according to our original parametric definition. If $\mathbf{x}$ satisfies this equation, then $x_1 = 3 + 2x_2 - 5x_3$ and so we may write

$$\begin{aligned} \mathbf{x} = (x_1, x_2, x_3) &= (3 + 2x_2 - 5x_3, x_2, x_3) \\ &= (3, 0, 0) + x_2(2, 1, 0) + x_3(-5, 0, 1). \end{aligned}$$

So, if we let $\mathbf{x}_0 = (3, 0, 0), \mathbf{u} = (2, 1, 0)$, and $\mathbf{v} = (-5, 0, 1)$, we see that $\mathbf{x} = \mathbf{x}_0 + x_2\mathbf{u} + x_3\mathbf{v}$, where x_2 and x_3 are arbitrary scalars. This is in accordance with our original definition of a plane in $\mathbb{R}^3$. ■

Finally, generalizing to n dimensions:

Definition: If $\mathbf{a} \in \mathbb{R}^n$ is a nonzero vector and $c \in \mathbb{R}$, then the equation $\mathbf{a} \cdot \mathbf{x} = c$ defines a *hyperplane in $\mathbb{R}^n$*. This means that the solution set has "dimension" $n - 1$, i.e., 1 less than the dimension of the ambient space $\mathbb{R}^n$.

Let's write an explicit formula for the general vector $\mathbf{x}$ satisfying this equation: If $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ and $a_1 \neq 0$, then we rewrite the equation $a_1x_1 + a_2x_2 + \cdots + a_nx_n = c$ to solve for x_1:

$$x_1 = \frac{1}{a_1}\left(c - a_2x_2 - \cdots - a_nx_n\right),$$

and so the general solution is of the form

$$\begin{aligned}\mathbf{x} = (x_1, \ldots, x_n) &= \left(\frac{1}{a_1}(c - a_2x_2 - \cdots - a_nx_n), x_2, \ldots, x_n\right) \\ &= \left(\frac{c}{a_1}, 0, \ldots, 0\right) + x_2\left(-\frac{a_2}{a_1}, 1, 0, \ldots, 0\right) + x_3\left(-\frac{a_3}{a_1}, 0, 1, \ldots, 0\right) \\ &+ \cdots + x_n\left(-\frac{a_n}{a_1}, 0, \ldots, 0, 1\right).\end{aligned}$$

(We leave it to the reader to write down the formula in the event that $a_1 = 0$.)

Example 3.8 *We give a parametric description of the line of intersection of the planes*

$$\begin{aligned}x_1 + 2x_2 - x_3 &= 2 \\ x_1 - x_2 + 2x_3 &= 5.\end{aligned}$$

Subtracting the first equation from the second yields $-3x_2+3x_3 = 3$*, or* $-x_2+x_3 = 1$*. Adding twice the latter equation to the first equation in the original system yields* $x_1+x_3 = 4$*. Thus, we can determine both* x_1 *and* x_2 *in terms of* x_3*:*

$$x_1 = 4 - x_3, \quad x_2 = -1 + x_3.$$

Then the general solution is of the form

$$\mathbf{x} = (x_1, x_2, x_3) = (4 - x_3, -1 + x_3, x_3) = (4, -1, 0) + x_3(-1, 1, 1).$$

The direction vector $(-1, 1, 1)$ *is orthogonal to* $\mathbf{a} = (1, 2, -1)$ *and* $\mathbf{b} = (1, -1, 2)$. ■

3.2.2 Flanigan and Kazdan, Calculus Two: Linear and Non-linear Functions

Let Y be a vector and $\mathcal{S}$ a linear subspace of $\mathbb{R}^n$. Projecting Y onto $\mathcal{S}$ means writing Y as $P + Q$, where P is a vector in $\mathcal{S}$ and Q is perpendicular to every vector in $\mathcal{S}$. We will show in the next theorem that this decomposition of Y into the sum $P + Q$ is unique. The vector P is called the orthogonal projection or simply the projection of Y onto $\mathcal{S}$.

Theorem: Let Y be a vector and $\mathcal{S}$ a linear subspace of $\mathbb{R}^n$. Suppose that $Y = P + Q$, where P is a vector in $\mathcal{S}$ and Q is orthogonal to every vector in $\mathcal{S}$. Then for any vector Z in $\mathcal{S}$ other than P, $\|Y - P\| < \|Y - Z\|$. The vector P is the unique vector in $\mathcal{S}$ such that $Y - P$ is orthogonal to every vector in $\mathcal{S}$.

> Proof: Let Z be some vector in $\mathcal{S}$ other than P. Since $\mathcal{S}$ is a linear subspace and P is in $\mathcal{S}$, the vector $P - Z$ is also in $\mathcal{S}$. Since the vector $Q = Y - P$ is orthogonal to every vector in $\mathcal{S}$, $Y - P$ and $P - Z$ are orthogonal. Thus, the Pythagorean relationship holds for $Y - P$ and $P - Z$:
>
> $$\|Y - P\|^2 + \|P - Z\|^2 = \|(Y - P) + (P - Z)\|^2 = \|Y - Z\|^2.$$
>
> Since P is not equal to Z, $\|P - Z\|^2 > 0$. It follows that $\|Y - P\|^2 < \|Y - Z\|^2$, or equivalently, $\|Y - P\| < \|Y - Z\|$ for every vector Z in $\mathcal{S}$ other than P. We have shown that if P satisfies the hypothesis of the theorem, then P is closer to Y than any other vector in $\mathcal{S}$. Therefore, P is the only vector that satisfies that hypothesis.

I interject here another proof of uniqueness. It is instructive, and more general, as it applies to any inner product. Recall the definition of the inner product, in (3.1), as, specifically for our context, the dot product for vectors in $\mathbb{R}^n$, and the properties listed there. The proof comes from Atanasiu and Mikusinski, Linear Algebra: Core Topics for the Second Course, 2023, p. 134.

Theorem: Let $\mathcal{U}$ be a subspace of an inner product space $\mathcal{V}$ and let $\mathbf{v} \in \mathcal{V}$. If an orthogonal projection of $\mathbf{v}$ on the subspace $\mathcal{U}$ exists, then it is unique.

> Proof: Assume that both $\mathbf{p}_1$ and $\mathbf{p}_2$ are orthogonal projections of $\mathbf{v}$ on the subspace $\mathcal{U}$, that is, $\langle \mathbf{v} - \mathbf{p}_1, \mathbf{u} \rangle = \langle \mathbf{v} - \mathbf{p}_2, \mathbf{u} \rangle = 0$ for every $\mathbf{u} \in \mathcal{U}$. Since $\mathbf{p}_1, \mathbf{p}_2 \in \mathcal{U}$, we have
>
> $$0 = \langle \mathbf{v} - \mathbf{p}_1, \mathbf{p}_2 \rangle = \langle \mathbf{v}, \mathbf{p}_2 \rangle - \langle \mathbf{p}_1, \mathbf{p}_2 \rangle$$
>
> and
>
> $$0 = \langle \mathbf{v} - \mathbf{p}_2, \mathbf{p}_2 \rangle = \langle \mathbf{v}, \mathbf{p}_2 \rangle - \langle \mathbf{p}_2, \mathbf{p}_2 \rangle = \langle \mathbf{v}, \mathbf{p}_2 \rangle - \|\mathbf{p}_2\|^2 .$$
>
> Consequently, $\langle \mathbf{p}_1, \mathbf{p}_2 \rangle = \|\mathbf{p}_2\|^2$. Similarly, we can show that $\langle \mathbf{p}_2, \mathbf{p}_1 \rangle = \|\mathbf{p}_1\|^2$. Hence
>
> $$\|\mathbf{p}_1 - \mathbf{p}_2\|^2 = \langle \mathbf{p}_1 - \mathbf{p}_2, \mathbf{p}_1 - \mathbf{p}_2 \rangle = \|\mathbf{p}_1\|^2 - \langle \mathbf{p}_1, \mathbf{p}_2 \rangle - \langle \mathbf{p}_2, \mathbf{p}_1 \rangle + \|\mathbf{p}_2\|^2 = 0,$$
>
> proving that $\mathbf{p}_1 = \mathbf{p}_2$.

The projection P of Y onto $\mathcal{S}$ is closer to Y than any other vector in $\mathcal{S}$. As such, we have:

Definition: The quantity $\|Y - P\| = \|Q\|$ is called the *distance between the point Y and linear subspace $\mathcal{S}$*.

Example 3.9 *Let us find the projection of the vector $Y = (-2, -6, -17)$ onto the plane $\mathcal{S}$ spanned by $X_1 = (1, 1, -2)$ and $X_2 = (1, -5, -4)$.*

This is conceptually illustrated in Figure 19, with the shown vectors not corresponding to the numeric values in Y, X_1, and X_2. It shows the projection of vector $\mathbf{v}$, denoted $P(\mathbf{v})$, onto the plane $\mathcal{S}$ spanned by $\mathbf{w}_1$ and $\mathbf{w}_2$.

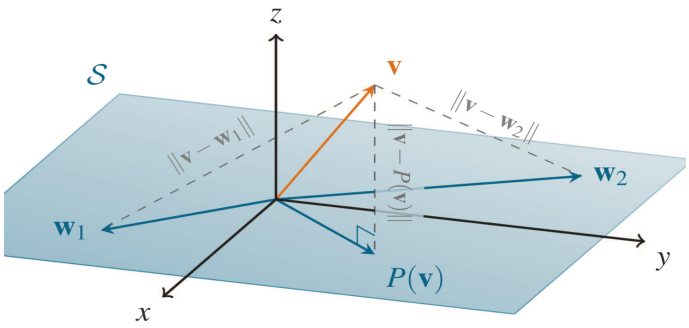


Figure 19: From Johnston, Advanced Linear and Matrix Algebra, 2021, p. 105.

We are looking for a vector P of the form $\alpha_1(1, 1, -2) + \alpha_2(1, -5, -4)$ such that the vector $Q = Y - P$ is orthogonal to every vector in $\mathcal{S}$. Note that if Q is orthogonal to both X_1 and

X_2, then Q is orthogonal to every linear combination of X_1 and X_2 (can you see why?), so Q is orthogonal to every vector in $\mathcal{S}$. Thus, we only need P to satisfy

$$\langle Y - P, X_1 \rangle = 0 \quad \text{and} \quad \langle Y - P, X_2 \rangle = 0.$$

Substituting in the values for Y, X_1, and X_2 and the expression for P, we obtain the following two equations in the unknowns α_1 and α_2:

$$\begin{aligned} \langle (-2,-6,-17) - \alpha_1(1,1,-2) - \alpha_2(1,-5,-4), (1,1,-2) \rangle &= 0, \\ \langle (-2,-6,-17) - \alpha_1(1,1,-2) - \alpha_2(1,-5,-4), (1,-5,-4) \rangle &= 0, \end{aligned}$$

which, after the various inner products are computed, become

$$6\alpha_1 + 4\alpha_2 = 26, \qquad 4\alpha_1 + 42\alpha_2 = 96.$$

This system of two equations is easily solved using the Gauss-Jordan method, with solution $\alpha_1 = 3$, $\alpha_2 = 2$. The projection of Y onto $\mathcal{S}$ is $P = 3(1,1,-2) + 2(1,-5,-4) = (5,-7,-14)$. The distance between Y and $\mathcal{S}$ is the norm of the vector

$$Q = Y - P = (-2,-6,-17) - (5,-7,-14) = (-7,1,-3),$$

which is $\sqrt{59}$. You should verify that Q is orthogonal to the vectors P, X_1, and X_2.

This illustrates the general method for projecting a vector Y onto a plane $\mathcal{S}$ in $\mathbb{R}^n$. If $\mathcal{S}$ is spanned by X_1 and X_2, we look for a vector P of the form $\alpha_1 X_1 + \alpha_2 X_2$ such that $Y - P$ is orthogonal to both X_1 and X_2. In terms of inner products,

$$\langle Y - \alpha_1 X_1 - \alpha_2 X_2, X_1 \rangle = 0 \quad \text{and} \quad \langle Y - \alpha_1 X_1 - \alpha_2 X_2, X_2 \rangle = 0.$$

Using the properties of inner products, these two equations may be rewritten as

$$\begin{aligned} \alpha_1 \langle X_1, X_1 \rangle + \alpha_2 \langle X_1, X_2 \rangle &= \langle Y, X_1 \rangle \\ \alpha_1 \langle X_1, X_2 \rangle + \alpha_2 \langle X_2, X_2 \rangle &= \langle Y, X_2 \rangle . \end{aligned} \tag{3.15}$$

Solve these two equations for the unknowns α_1 and α_2. Take the corresponding linear combination of X_1 and X_2 to obtain the projection P. The distance between Y and $\mathcal{S}$ is $\|Y - P\|$. It is interesting to note that, even though the vectors Y, X_1, and X_2 are in $\mathbb{R}^n$, we always obtain two equations in two unknowns when finding the projection of a vector onto a plane, whatever the value of n.

The method just outlined extends naturally to projections onto k-dimensional linear subspaces. If $\mathcal{S}$ is spanned by the vectors $X_1, \ldots, X_k$, then we want the vector P to satisfy $\langle Y - P, X_i \rangle = 0$ for $i = 1, \ldots, k$. Substitute $\alpha_1 X_1 + \cdots + \alpha_k X_k$ for P to obtain k equations in the k unknowns $\alpha_1, \ldots, \alpha_k$. Solve the equations and then compute P. The projection of Y onto $\mathcal{S}$ is P, and the distance between Y and $\mathcal{S}$ is $\|Y - P\|$. ■

Remark 1: We will later see a name for the 2×2 matrix implied in (3.15), and its $k \times k$ extension discussed in the previous paragraph: This is called the Gram matrix, as given in (3.39). Indeed, with $\mathbf{K}$ denoting the Gram matrix of inner products, $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_k)'$, and $\mathbf{c} = (c_1, \ldots, c_k)'$, with $c_i = \langle Y, X_i \rangle$, we can express (3.15) as $\mathbf{K}\boldsymbol{\alpha} = \mathbf{c}$. If $\mathbf{K}$ is full rank, which is equivalent to $\{X_i\}$ being linearly independent (with $k = 2$, this means, X_1 and X_2 are not collinear), then there is a unique solution for $\boldsymbol{\alpha}$. If the $\{X_i\}$ are orthogonal, or orthonormal,

as induced by, e.g., Gram-Schmidt, then the solution is very easy to express. We address this in §3.4.3 below.

Remark 2: Another way of posing the task in Example 3.9 is to give a point on the plane $\mathcal{S}$ (for which we can take X_1, or X_2, or any nonzero linear combination of them) and the normal vector N, to the plane. Normal N can be computed using the cross product from (3.13) based on X_1 and X_2, giving $(-14, 2, -6)$, which is parallel to $N = (-7, 1, -3)$.

We now change notation, because we wish to use the method in Example 3.13, which poses a similar question, but gives, as we just computed, a point on the plane, called P, and take P to be $X_1 = (1, 1, -2)$; the point Y, which we rename to $Q = (-2, -6, -17)$; and the normal vector N, as just computed. Plugging these into (3.22), we get the distance from Q to the plane:

$$\frac{|(Q-P)\cdot N|}{\|N\|} = \frac{|((-2,-6,-17)-(1,1,-2))\cdot(-7,1,-3)|}{\|(-7,1,-3)\|} = \frac{|59|}{\sqrt{59}} = \sqrt{59},$$

as before.

Suppose one is interested in finding the "projection" of a vector Y onto an affine subspace $\mathcal{A}$ and the "distance" from Y to $\mathcal{A}$. Here is the issue described more precisely. We want to find P and Q such that $Y = P + Q$, $P \in \mathcal{A}$, and Q is perpendicular to the difference of any two members of $\mathcal{A}$. It develops that this can always be done, and in only one way. The vector P, thus uniquely determined, is the projection of Y onto $\mathcal{A}$, and $\|Q\|$ is the distance from Y to $\mathcal{A}$.

Write $\mathcal{A} = \mathcal{S} + Z$ for some Z. Write $Y - Z = R + Q$, where R is the projection of $Y - Z$ onto the linear subspace $\mathcal{S}$. Then $Y = P + Q$, where $P = R + Z$. Now let us check that P is the projection of Y onto the affine subspace $\mathcal{A}$ and that, therefore, $\|Q\|$ is the distance from Y to $\mathcal{A}$. It is clear that $P \in \mathcal{A}$ since it is the sum of Z and a member of $\mathcal{S}$. That Q is perpendicular to the difference of any two members of $\mathcal{A}$ follows from the facts that Q is perpendicular to every member of $\mathcal{S}$ and that the difference of any two members of $\mathcal{A}$ is a member of $\mathcal{S}$. The drawing of an accurate picture for the following example can be of help in grasping these ideas.

Example 3.10 *Let us calculate the projection of* $(3, 5)$ *onto the line* $\mathcal{S} + (-1, 3)$ *where* $\mathcal{S}$ *is the one-dimensional linear subspace spanned by* $(3, 4)$*. We subtract* $(-1, 3)$ *from* $(3, 5)$ *and find the projection of the difference* $(4, 2)$ *onto* $(3, 4)$*:*

$$\frac{\langle (3,4),(4,2)\rangle}{\|(3,4)\|^2}(3,4) = \left(\frac{12}{5}, \frac{16}{5}\right).$$

Thus, the projection of $(3, 5)$ *onto* $\mathcal{S} + (-1, 3)$ *is* $\left(\frac{12}{5}, \frac{16}{5}\right) + (-1, 3) = \left(\frac{7}{5}, \frac{31}{5}\right)$*. To get the distance from* $(3, 5)$ *to* $\mathcal{S} + (-1, 3)$ *we subtract its projection from it and take the norm:*

$$\left\| \left(3 - \frac{7}{5}, 5 - \frac{31}{5}\right) \right\| = \sqrt{4} = 2. \qquad \blacksquare$$

3.2.3 Lang, Calculus of Several Variables

We define a *located vector* to be an ordered pair of points that we write $\overrightarrow{AB}$. (This is not a product.) We visualize this as an arrow between A and B. We call A the beginning point and B the end point of the located vector. The difference, $B - A$, is defined by writing $B = A + (B - A)$. Let $\overrightarrow{AB}$ and $\overrightarrow{CD}$ be two located vectors. We shall say that they are equivalent if $B - A = D - C$. Every located vector $\overrightarrow{AB}$ is equivalent to one whose beginning point is the origin, because $\overrightarrow{AB}$ is equivalent to $\overrightarrow{O(B-A)}$ (see the left panel of Figure 20). Clearly this is the only located vector whose beginning point is the origin and which is equivalent to $\overrightarrow{AB}$. If you visualize the parallelogram law in the plane, then it is clear that equivalence of two located vectors can be interpreted geometrically by saying that the lengths of the line segments determined by the pair of points are equal, and that the "directions" in which they point are the same. Figure 20 shows the located vectors $\overrightarrow{O(B-A)}$, $\overrightarrow{AB}$, and $\overrightarrow{O(A-B)}$, $\overrightarrow{BA}$.

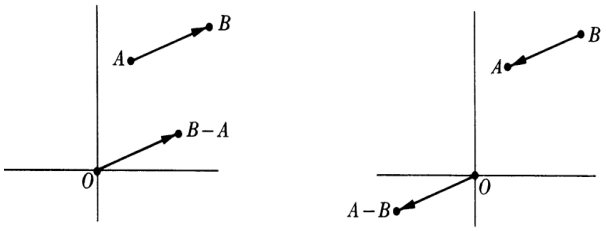


Figure 20: Various located vectors. From Lang, p. 12

Given a located vector $\overrightarrow{OC}$ whose beginning point is the origin, we shall say that it is located at the origin. Given any located vector $\overrightarrow{AB}$, we shall say that it is located at A. A located vector at the origin is entirely determined by its end point. In view of this, we shall call an n-tuple either a point or a vector, depending on the interpretation which we have in mind. Two located vectors $\overrightarrow{AB}$ and $\overrightarrow{PQ}$ are said to be parallel if there is a number $c \neq 0$ such that $B - A = c(Q - P)$. They are said to have the same direction if there is a number $c > 0$ such that $B - A = c(Q - P)$, and have opposite direction if there is a number $c < 0$ such that

$$B - A = c(Q - P).$$

Instead of writing $A \cdot A$ for the scalar product of a vector with itself, it will be convenient to write also A^2. (This is the only instance when we allow ourselves such a notation. Thus A^3 has no meaning.) As an exercise, verify the following identities using the properties listed after (3.1):

$$(A + B)^2 = A^2 + 2A \cdot B + B^2, \qquad (A - B)^2 = A^2 - 2A \cdot B + B^2. \tag{3.16}$$

As previously stated, for A and B be two points in $\mathbb{R}^n$, we define the distance between them to be

$$\|A - B\| = \sqrt{(A - B) \cdot (A - B)}.$$

This definition coincides with our geometric intuition when A, B are points in the plane; see the left panel of Figure 20. It is the same thing as the length of the located vector $\overrightarrow{AB}$ or the located vector $\overrightarrow{BA}$.

We shall say that a vector E is a unit vector if $\|E\| = 1$. Given any vector A, let $a = \|A\|$. If $a \neq 0$, then A/a is a unit vector, because

$$\left\| \frac{1}{a} A \right\| = \frac{1}{a} a = 1.$$

We say that two vectors A, B (neither of which is O) have the same direction if there is a number $c > 0$ such that $cA = B$. In view of this definition, the vector $A/\|A\|$ is a unit vector in the direction of A (provided $A \neq O$).

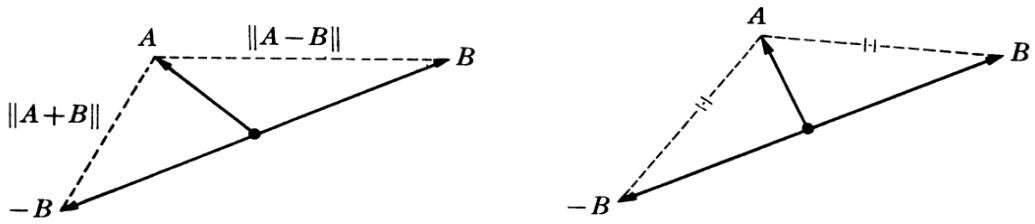


Figure 21: From Lang, p. 24

As stated above in (3.6), we define two (compatible) vectors A and B to be perpendicular, or orthogonal, if $A \cdot B = 0$. The method of proof from Lang is different than that above.

Given A, B in the plane, the condition that $\|A + B\| = \|A - B\|$ (illustrated in Figure 21) coincides with the geometric property that A should be perpendicular to B. To prove

$$\|A + B\| = \|A - B\| \text{ if and only if } A \cdot B = 0, \tag{3.17}$$

use (3.16) to write

$$\begin{aligned} \|A + B\| = \|A - B\| &\Leftrightarrow \|A + B\|^2 = \|A - B\|^2 \\ &\Leftrightarrow A^2 + 2A \cdot B + B^2 = A^2 - 2A \cdot B + B^2 \\ &\Leftrightarrow 4A \cdot B = 0 \Leftrightarrow A \cdot B = 0. \end{aligned}$$

Theorem (The General Pythagoras Theorem): If A and B are perpendicular, then

$$\|A + B\|^2 = \|A\|^2 + \|B\|^2. \tag{3.18}$$

Proof: Use the definitions, namely

$$\|A + B\|^2 = (A + B) \cdot (A + B) = A^2 + 2A \cdot B + B^2 = \|A\|^2 + \|B\|^2,$$

because $A \cdot B = 0$, and $A \cdot A = \|A\|^2$, $B \cdot B = \|B\|^2$ by definition.

Note: If A is perpendicular to B, and x is any number, then A is also perpendicular to xB because $A \cdot xB = xA \cdot B = 0$.

We shall now use the notion of perpendicularity to derive the notion of projection. Let A, B be two vectors and $B \neq O$. Let P be the point on the line through $\overrightarrow{OB}$ such that $\overrightarrow{PA}$ is perpendicular to $\overrightarrow{OB}$, as shown in Figure 22. We can write $P = cB$ for some number c. We want to find this number c explicitly in terms of A and B. The condition $\overrightarrow{PA} \perp \overrightarrow{OB}$ means

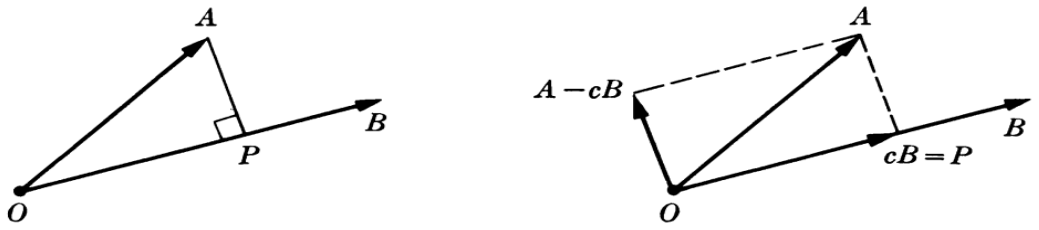


Figure 22: From Lang, p. 25

that $A - P$ is perpendicular to B, and since $P = cB$ this means that $(A - cB) \cdot B = 0$; in other words, $A \cdot B - cB \cdot B = 0$. We can solve for c, and we find $A \cdot B = cB \cdot B$, so that

$$c = \frac{A \cdot B}{B \cdot B}.$$

Conversely, if we take this value for c, and then use distributivity, dotting $A - cB$ with B yields 0, so that $A - cB$ is perpendicular to B. Hence we have seen that there is a unique number c such that $A - cB$ is perpendicular to B, and c is given by the above formula. We define:

Definition: The *component of A along B* is the number $c = (A \cdot B)/(B \cdot B)$. The *projection of A along B* is the vector cB.

Our construction gives an immediate geometric interpretation for the scalar product. Namely, assume $A \neq O$ and consider the angle θ between A and B, using the left panel of Figure 22. Then, from plane geometry, we see that

$$\cos\theta = \frac{c\|B\|}{\|A\|},$$

or, substituting the value for c obtained above,

$$A \cdot B = \|A\|\|B\| \cos\theta \quad \text{and} \quad \cos\theta = \frac{A \cdot B}{\|A\|\|B\|}. \tag{3.19}$$

Using this construction, we will prove two fundamental inequalities. First observe: If $E_i = (0, \ldots, 0, 1, 0, \ldots, 0)$ is the ith unit vector of $\mathbb{R}^n$, and $A = (a_1, \ldots, a_n)$, then $A \cdot E_i = a_i$ is the ith component of A, i.e., the component of A along E_i. We have

$$|a_i| = \sqrt{a_i^2} \leqq \sqrt{a_1^2 + \cdots + a_n^2} = \|A\|,$$

so that the absolute value of each component of A is at most equal to the length of A. More generally, let E be any unit vector (a vector of norm 1). Let c be the component of A along E, which reduces to $c = A \cdot E$. Then $A - cE$ is perpendicular to E, $A = A - cE + cE$, and $A - cE$ is also perpendicular to cE. Thus, by Pythagoras,

$$\|A\|^2 = \|A - cE\|^2 + \|cE\|^2 = \|A - cE\|^2 + c^2,$$

and we have the inequality $c^2 \leqq \|A\|^2$, i.e., $|c| \leqq \|A\|$.

We now generalize this inequality to a dot product $A \cdot B$ when B is not necessarily a unit vector, yielding the Cauchy-Schwarz inequality (1.9), where it was proven in a different way.

Theorem (The (Cauchy-)Schwarz Inequality): Let A, B be two vectors in $\mathbb{R}^n$. Then

$$|A \cdot B| \leqq \|A\|\|B\|. \tag{3.20}$$

Proof: If $B = O$, then both sides of the inequality are equal to 0, and so our assertion is obvious. Suppose that $B \neq O$. Let c be the component of A along B, so $c = (A \cdot B)/(B \cdot B)$. We write $A = A - cB + cB$, and, by Pythagoras,

$$\|A\|^2 = \|A - cB\|^2 + \|cB\|^2 = \|A - cB\|^2 + c^2\|B\|^2.$$

Hence $c^2\|B\|^2 \leqq \|A\|^2$. But

$$c^2\|B\|^2 = \frac{(A \cdot B)^2}{(B \cdot B)^2}\|B\|^2 = \frac{|A \cdot B|^2}{\|B\|^4}\|B\|^2 = \frac{|A \cdot B|^2}{\|B\|^2},$$

or

$$\frac{|A \cdot B|^2}{\|B\|^2} \leqq \|A\|^2.$$

Multiply by $\|B\|^2 > 0$ and take the square root to conclude the proof.

In view of the (Cauchy-)Schwarz inequality, we see that, for vectors A, B in n-space, the number $(A \cdot B)/(\|A\|\|B\|)$ has absolute value bounded by 1. Consequently,

$$-1 \leqq \frac{A \cdot B}{\|A\|\|B\|} \leqq 1,$$

and there exists a unique angle θ such that $0 \leqq \theta \leqq \pi$, and such that

$$\cos\theta = \frac{A \cdot B}{\|A\|\|B\|}.$$

Definition: We define this angle to be the angle between A and B.

As in (1.10), we use Cauchy-Schwarz to prove:

Theorem (The Triangle Inequality): Let A, B be vectors. Then

$$\|A + B\| \leqq \|A\| + \|B\|. \tag{3.21}$$

Proof: Both sides of this inequality are positive or 0. Hence it will suffice to prove that their squares satisfy the desired inequality, in other words,

$$(A + B) \cdot (A + B) \leqq (\|A\| + \|B\|)^2.$$

To do this, use $(A + B) \cdot (A + B) = A \cdot A + 2A \cdot B + B \cdot B$ from (3.16), which, as just demonstrated, satisfies the inequality $\leqq \|A\|^2 + 2\|A\|\|B\| + \|B\|^2$, the rhs of which is none other than $(\|A\| + \|B\|)^2$.

The name triangle inequality comes from the following: If we draw a triangle as in the left panel of Figure 23, then the triangle inequality expresses the fact that the length of one side is bounded by the sum of the lengths of the other two sides.

We now revisit the material on planes and their mathematical representations, giving further, different, and useful examples from Lang.

Let P be a point in 3-space and consider a located vector $\overrightarrow{ON}$. We define the plane passing through P perpendicular to $\overrightarrow{ON}$ to be the collection of all points X such that the

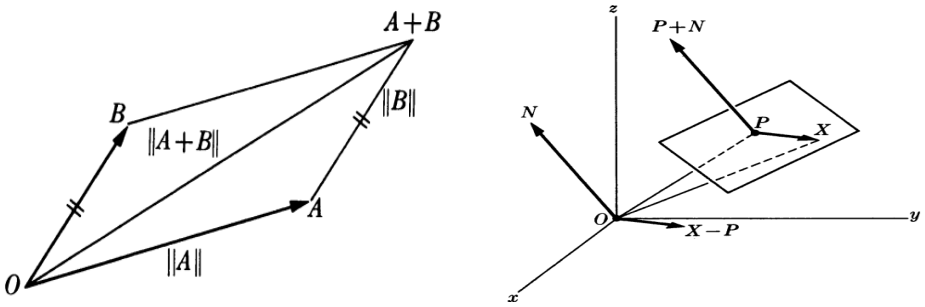


Figure 23: From Lang, p. 30; and 36

located vector $\overrightarrow{PX}$ is perpendicular to $\overrightarrow{ON}$. According to our definitions, this amounts to the condition $(X-P)\cdot N = 0$, which can also be written as $X\cdot N = P\cdot N$. We shall also say that this plane is the one perpendicular to N, and consists of all vectors X such that $X-P$ is perpendicular to N. The right panel of Figure 23 shows a typical situation in 3-space.

Instead of saying that N is perpendicular to the plane, one also says that N is normal to the plane. Let t be a number $\neq 0$. Then the set of points X such that $(X-P)\cdot N = 0$ coincides with the set of points X such that $(X-P)\cdot tN = 0$. Thus we may say that our plane is the plane passing through P and perpendicular to the line in the direction of N. To find the equation of the plane, we could use any vector tN (with $t\neq 0$) instead of N.

Example 3.11 *Let $Q = (1,1,1)$, $P = (1,-1,2)$, and $N = (1,2,3)$. We wish to find the point of intersection of the line through P in the direction of N, and the plane through Q perpendicular to N. The parametric representation of the line through P in the direction of N is $X = P+tN$. The equation of the plane through Q perpendicular to N is $(X-Q)\cdot N = 0$; and is visualized in Figure 24.*

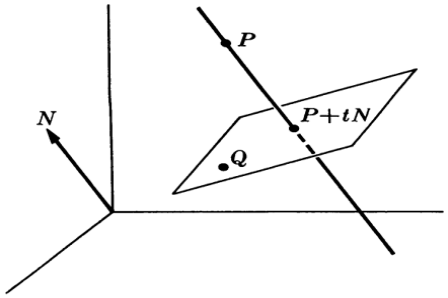


Figure 24: From Lang, p. 39

We must find t such that $X = P+tN$ also satisfies $(X-Q)\cdot N = 0$; that is, $(P+tN-Q)\cdot N = 0$, or, after using the rules of the dot product, $(P-Q)\cdot N + tN\cdot N = 0$. Solving for t yields

$$t = \frac{(Q-P)\cdot N}{N\cdot N} = \frac{1}{14}.$$

Thus,

$$P+tN = (1,-1,2)+\frac{1}{14}(1,2,3) = \left(\frac{15}{14},-\frac{12}{14},\frac{31}{14}\right)$$

is the desired point of intersection. ■

Example 3.12 *Find the equation of the plane passing through the three points*

$$P_1 = (1, 2, -1), \quad P_2 = (-1, 1, 4), \quad P_3 = (1, 3, -2).$$

We find a vector N perpendicular to $\overrightarrow{P_1P_2}$ and $\overrightarrow{P_1P_3}$, or in other words, perpendicular to $P_2 - P_1$ and $P_3 - P_1$. We have $P_2 - P_1 = (-2, -1, +5)$ and $P_3 - P_1 = (0, 1, -1)$. Let $N = (a, b, c)$. We must solve $N \cdot (P_2 - P_1) = 0$ and $N \cdot (P_3 - P_1) = 0$, or

$$-2a - b + 5c = 0, \qquad b - c = 0.$$

We take $b = c = 1$ and solve for $a = 2$. Then $N = (2, 1, 1)$ satisfies our requirements. The plane perpendicular to N, passing through P_1 is the desired plane. Its equation is therefore $X \cdot N = P_1 \cdot N$, that is, $2x + y + z = 2 + 2 - 1 = 3$. ■

Example 3.13 *Consider a plane defined by the equation $(X - P) \cdot N = 0$, and let Q be an arbitrary point. We wish to find a formula for the distance between Q and the plane. By this we mean the length of the segment from Q to the point of intersection of the perpendicular line to the plane through Q, as shown in Figure 25. We let Q' be this point of intersection.*

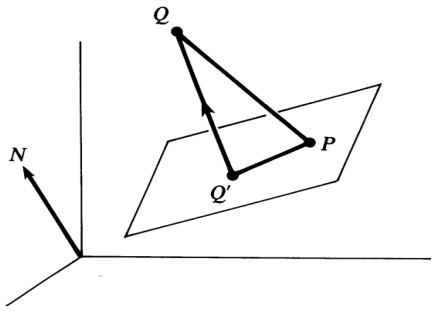


Figure 25: From Lang, p. 41

From the geometry, we have: length of the segment $\overline{QQ'}$ = length of the projection of $\overline{QP}$ on $\overline{QQ'}$. We can express the length of this projection in terms of the dot product as follows. A unit vector in the direction of N, which is perpendicular to the plane, is given by $N/\|N\|$. Then

$$\begin{aligned} &\text{length of the projection of } \overline{QP} \text{ on } \overline{QQ'} \\ &= \text{ norm of the projection of } Q - P \text{ on } N/\|N\| \\ &= \left| (Q - P) \cdot \frac{N}{\|N\|} \right| = \frac{|(Q - P) \cdot N|}{\|N\|}, \end{aligned} \tag{3.22}$$

this being the distance between Q and the plane. ■

See also Remark 2 of Example 3.9.

3.3 Determinants and the Cross Product

We also assume the reader has basic familiarity with matrix algebra, e.g., multiplication, transpose, symmetry, determinant, and inverse. This will be relevant in this subsection, and notably also in §4.6 and §5.6.

In the remainder of this subsection, we look at the cross product, basing our presentation on (an augmented version of that in) Shifrin, Multivariable Mathematics, 2005, and (mostly verbatim) Lang, Calculus of Several Variables, 3rd ed., 1987. We begin with the former.

Let $\mathbf{x}$ and $\mathbf{y}$ be vectors in $\mathbb{R}^2$ and consider the parallelogram $\mathcal{P}$ they span. The area of $\mathcal{P}$ is nonzero as long as $\mathbf{x}$ and $\mathbf{y}$ are not collinear. We want to express the area of $\mathcal{P}$ in terms of the coordinates of $\mathbf{x}$ and $\mathbf{y}$. First notice that the area of the parallelogram pictured in the left panel of Figure 26 is the same as the area of the rectangle obtained by moving the shaded triangle from the right side to the left. This rectangle has area bh, where $b = \|\mathbf{x}\|$ is the base and $h = \|\mathbf{y}\| \sin\theta$ is the height.

We could calculate $\sin\theta$ from (3.19), namely $\cos\theta = (\mathbf{x}\cdot\mathbf{y})/\|\mathbf{x}\|\|\mathbf{y}\|$, but instead we note from the middle panel of Figure 26, and recalling (3.5), or (3.14), or (3.19),

$$\text{area}(\mathcal{P}) = bh = \|\mathbf{x}\|\|\mathbf{y}\| \sin\theta = \|\mathbf{x}\|\|\mathbf{y}\| \cos\left(\frac{\pi}{2} - \theta\right) = \rho(\mathbf{x}) \cdot \mathbf{y}, \tag{3.23}$$

where $\rho(\mathbf{x})$ is the vector obtained by rotating $\mathbf{x}$ an angle $\pi/2$ counterclockwise.

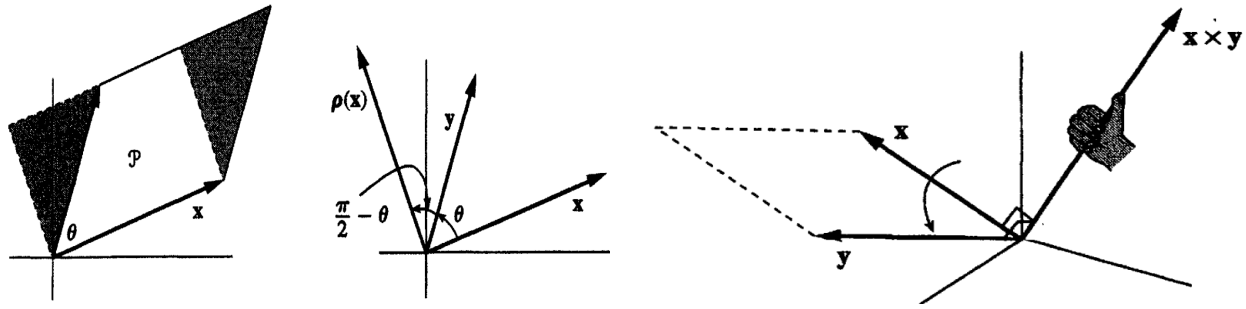


Figure 26: From Shifrin, p. 44 and p. 48. In the middle panel, imagine a perpendicular line, dropped from the end of vector $\mathbf{y}$ onto $\rho(\mathbf{x})$, which is the projection of $\mathbf{y}$ onto $\rho(\mathbf{x})$. The length of this projection is, from the usual determination of cosine from an angle drawn in a cirlce, $\|\mathbf{y}\| \cos\left(\frac{\pi}{2} - \theta\right)$. In the third panel: If you curl the fingers of your right hand from $\mathbf{x}$ toward $\mathbf{y}$, your thumb points in the direction of $\mathbf{x} \times \mathbf{y}$.

Note the known identities (e.g., Wikipedia) also from (2.50) and (2.52):

$$\cos\theta = \sin\left(\frac{\pi}{2} - \theta\right), \quad \sin\theta = \cos\left(\frac{\pi}{2} - \theta\right), \quad \sin\left(\theta \pm \frac{\pi}{2}\right) = \pm\cos\theta, \quad \cos\left(\theta \pm \frac{\pi}{2}\right) = \mp\sin\theta. \tag{3.24}$$

If $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$, then we have $\rho(\mathbf{x}) \cdot \mathbf{y} = \begin{bmatrix} -x_2 \\ x_1 \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$, i.e.,

$$\|\mathbf{x}\|\|\mathbf{y}\| \sin\theta = \text{area}(\mathcal{P}) = \rho(\mathbf{x}) \cdot \mathbf{y} = x_1 y_2 - x_2 y_1 = \text{Det}(\mathbf{x}, \mathbf{y}). \tag{3.25}$$

If $\mathbf{x} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} 4 \\ 3 \end{bmatrix}$, then the area of the parallelogram spanned by $\mathbf{x}$ and $\mathbf{y}$ is $x_1 y_2 - x_2 y_1 = 3 \cdot 3 - 1 \cdot 4 = 5$. On the other hand, if we interchange the two, letting $\mathbf{x} = \begin{bmatrix} 4 \\ 3 \end{bmatrix}$

and $\mathbf{y} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$, then we get $x_1y_2 - x_2y_1 = 4 \cdot 1 - 3 \cdot 3 = -5$. Certainly the parallelogram hasn't changed; nor does it make sense to have negative area. What is the explanation? In deriving our formula for the area above, we assumed $0 < \theta < \pi$; but if we must turn clockwise to get from $\mathbf{x}$ to $\mathbf{y}$, this means that θ is negative, resulting in a sign discrepancy in the area calculation. So we should amend our earlier result. We define the signed area of the parallelogram $\mathcal{P}$ to be the area of $\mathcal{P}$ when one turns counterclockwise from $\mathbf{x}$ to $\mathbf{y}$ and to be negative the area of $\mathcal{P}$ when one turns clockwise from $\mathbf{x}$ to $\mathbf{y}$. Then we have

$$\text{signed area } (\mathcal{P}) = x_1y_2 - x_2y_1.$$

Definition: Given two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^3$, define a vector, called their *cross product*, by

$$(x_2y_3 - x_3y_2, \;\; x_3y_1 - x_1y_3, \;\; x_1y_2 - x_2y_1) \tag{3.26}$$

or, with $\mathbf{e}_i$ the ith unit vector in $\mathbb{R}^3$ and a "formal"[24] use of the determinant,

$$\mathbf{x} \times \mathbf{y} = (x_2y_3 - x_3y_2)\,\mathbf{e}_1 + (x_3y_1 - x_1y_3)\,\mathbf{e}_2 + (x_1y_2 - x_2y_1)\,\mathbf{e}_3 \tag{3.27}$$

$$= \begin{vmatrix} \mathbf{e}_1 & x_1 & y_1 \\ \mathbf{e}_2 & x_2 & y_2 \\ \mathbf{e}_3 & x_3 & y_3 \end{vmatrix}. \tag{3.28}$$

Let $\mathbf{z} = (z_1, z_2, z_3)' = \sum z_i\mathbf{e}_i$, and note from (3.27) that

$$\mathbf{z} \cdot (\mathbf{x} \times \mathbf{y}) = \text{Det}(\mathbf{z}, \mathbf{x}, \mathbf{y}). \tag{3.29}$$

The geometric interpretation of the cross product is indicated in right panel of Figure 26, and given as follows. The cross product $\mathbf{x} \times \mathbf{y}$ of two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^3$ is orthogonal to both $\mathbf{x}$ and $\mathbf{y}$, and

$$\text{area}(\mathcal{P}) = \|\mathbf{x} \times \mathbf{y}\|. \tag{3.30}$$

Moreover, when $\mathbf{x}$ and $\mathbf{y}$ are nonparallel, the vectors $\mathbf{x}$, $\mathbf{y}$, $\mathbf{x} \times \mathbf{y}$ determine a parallelepiped of positive signed volume. To see this, note that the orthogonality is a consequence of (3.29) and properties of the determinant. In particular, $\mathbf{x} \cdot (\mathbf{x} \times \mathbf{y}) = \text{Det}(\mathbf{x}, \mathbf{x}, \mathbf{y}) = \mathbf{0}$. From the interpretation of determinant as the volume of a parallelepiped, $\text{Det}(\mathbf{x}, \mathbf{y}, \mathbf{x} \times \mathbf{y})$ is the signed volume of the parallelepiped spanned by $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{x} \times \mathbf{y}$. As $\mathbf{x} \times \mathbf{y}$ is orthogonal to the plane spanned by $\mathbf{x}$ and $\mathbf{y}$, that volume is the product of the area of $\mathcal{P}$ and $\|\mathbf{x} \times \mathbf{y}\|$. On the other hand, again from (3.29), simply substituting $\mathbf{x} \times \mathbf{y}$ in place of $\mathbf{z}$,

$$\text{Det}(\mathbf{x}, \mathbf{y}, \mathbf{x} \times \mathbf{y}) = \text{Det}(\mathbf{x} \times \mathbf{y}, \mathbf{x}, \mathbf{y}) = (\mathbf{x} \times \mathbf{y}) \cdot (\mathbf{x} \times \mathbf{y}) = \|\mathbf{x} \times \mathbf{y}\|^2. \tag{3.31}$$

Setting the two expressions equal yields $\|\mathbf{x} \times \mathbf{y}\| = \text{area}(\mathcal{P})$. When $\mathbf{x}$ and $\mathbf{y}$ are nonparallel, $\text{area}(\mathcal{P}) > 0$, or $\|\mathbf{x} \times \mathbf{y}\|^2 > 0$, so the three vectors span a parallelepiped of positive signed volume, as desired.

[24] According to Wikipedia, a formal calculation, or formal operation, is a calculation that is systematic but without a rigorous justification. It involves manipulating symbols in an expression using a generic substitution without proving that the necessary conditions hold.

Example 3.14 *We can use the cross product to find the equation of the subspace $\mathcal{P}$ spanned by the vectors* $u = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$ *and* $v = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$. *For the normal vector to P is*

$$\mathbf{A} = \mathbf{u} \times \mathbf{v} = \begin{vmatrix} \mathbf{e}_1 & 1 & 1 \\ \mathbf{e}_2 & 1 & 2 \\ \mathbf{e}_3 & -1 & 1 \end{vmatrix} = \begin{bmatrix} 3 \\ -2 \\ 1 \end{bmatrix}$$

and so

$$\mathcal{P} = \left\{ \mathbf{x} \in \mathbb{R}^3 : \mathbf{A} \cdot \mathbf{x} = 0 \right\} = \left\{ \mathbf{x} \in \mathbb{R}^3 : 3x_1 - 2x_2 + x_3 = 0 \right\}.$$

Moreover, the affine plane $\mathcal{P}_1$ parallel to $\mathcal{P}$ and passing through the point $x_0 = \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix}$ *is given by*

$$\begin{aligned} \mathcal{P}_1 &= \left\{ \mathbf{x} \in \mathbb{R}^3 : \mathbf{A} \cdot (\mathbf{x} - \mathbf{x}_0) = 0 \right\} = \left\{ \mathbf{x} \in \mathbb{R}^3 : \mathbf{A} \cdot \mathbf{x} = \mathbf{A} \cdot \mathbf{x}_0 \right\} \\ &= \left\{ \mathbf{x} \in \mathbb{R}^3 : 3x_1 - 2x_2 + x_3 = 7 \right\}. \qquad \blacksquare \end{aligned}$$

We now turn to the presentation in Lang.

Definition: Let $A = (a_1, a_2, a_3)$ and $B = (b_1, b_2, b_3)$ be two vectors in 3-space. We define their *cross product* to be the vector

$$A \times B = (a_2 b_3 - a_3 b_2,\ \ a_3 b_1 - a_1 b_3,\ \ a_1 b_2 - a_2 b_1), \tag{3.32}$$

which is symbolically (or "formally") the determinant

$$A \times B = \begin{vmatrix} E_1 & E_2 & E_3 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix}. \tag{3.33}$$

(Note (3.27) is the same as (3.32); and (3.28) agrees with (3.33).) Indeed, the rhs of (3.33) is, upon expansion, $E_1 (a_2 b_3 - a_3 b_2) - E_2 (a_1 b_3 - a_3 b_1) + E_3 (a_1 b_2 - a_2 b_1)$, which agrees with the definition of $A \times B$ in (3.32).

Theorem: The following results on the cross product can be verified directly from (3.32):

CP 1. $A \times B = -(B \times A)$.
CP 2. $A \times (B + C) = (A \times B) + (A \times C)$, and $(B + C) \times A = B \times A + C \times A$.
CP 3. For any number a, we have

$$(aA) \times B = a(A \times B) = A \times (aB).$$

CP 4. $(A \times B) \times C = (A \cdot C)B - (B \cdot C)A$.
CP 5. $A \times B$ is perpendicular to both A and B.
CP 6. $(A \times B)^2 = (A \times B) \cdot (A \times B) = (A \cdot A)(B \cdot B) - (A \cdot B)^2$.

The first three also follow immediately from results on determinants. For CP5,

$$A \cdot (A \times B) = a_1 (a_2 b_3 - a_3 b_2) + a_2 (a_3 b_1 - a_1 b_3) + a_3 (a_1 b_2 - a_2 b_1) = 0,$$

because all terms cancel. Similarly for $B \cdot (A \times B)$. The vector $A \times B$ is perpendicular to the plane spanned by A and B. So is $B \times A$, but $B \times A$ points in the opposite direction.

For CP6, the first equality is just Lang's notation mentioned just before (3.16). For the second equality, simple algebra based on definitions yields

$$(A \times B) \cdot (A \times B) = (a_2 b_3 - a_3 b_2)^2 + (a_3 b_1 - a_1 b_3)^2 + (a_1 b_2 - a_2 b_1)^2 ,$$
$$(A \cdot A)(B \cdot B) - (A \cdot B)^2 = \left(a_1^2 + a_2^2 + a_3^2\right)\left(b_1^2 + b_2^2 + b_3^2\right) - (a_1 b_1 + a_2 b_2 + a_3 b_3)^2 .$$

Expanding everything out and comparing confirms relation CP6.

From (3.5), or (3.14), or (3.19), CP6 is $\|A \times B\|^2 = \|A\|^2 \|B\|^2 - \|A\|^2 \|B\|^2 \cos^2 \theta$, where θ is the angle between A and B. Hence, we obtain $\|A \times B\|^2 = \|A\|^2 \|B\|^2 \sin^2 \theta$, or

$$\|A \times B\| = \|A\| \|B\| |\sin \theta|, \tag{3.34}$$

(which, of course, agrees with the juxtaposition of (3.23) and (3.30) in our first presentation above).

Lang concludes by showing what we saw above, namely (3.30): Identity (3.34) can be used to make another interpretation of the cross product. Indeed, we see that $\|A \times B\|$ is the area of the parallelogram spanned by A and B. If we consider the plane containing the located vectors $\overrightarrow{OA}$ and $\overrightarrow{OB}$, then the picture looks like that in Figure 27, and our assertion amounts simply to the statement that the area of a parallelogram is equal to the base times the altitude.

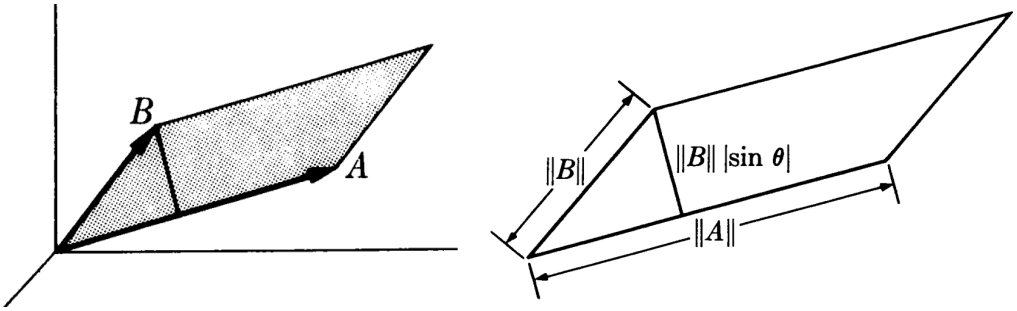


Figure 27: From Lang, p. 46 and p. 47

3.4 More Advanced Linear Algebra: Projection and Least Squares

This subsection is taken from the advanced undergraduate linear algebra book of Olver and Shakiban (Applied Linear Algebra, 2nd ed., 2018), and will presuppose an acquaintance with the numerous *basic* concepts from an introductory course in linear algebra, e.g., definitions of a vector space, subspaces, an independent set of vectors, the span of a set of vectors, a basis of a subspace, and the kernel and image of a matrix. We review the kernel and image below. We do not require previous knowledge of concepts such as change of basis, LDU, QR, and spectral (i.e., eigenvalue, eigenvector) factorizations or decompositions, Gram-Schmidt, or complex vector spaces. The Gram-Schmidt method and the resulting QR decomposition will be developed herein.

The image of an $m \times n$ matrix A is the subspace $\operatorname{img} A \subset \mathbb{R}^m$ spanned by its columns. The kernel of A is the subspace $\ker A \subset \mathbb{R}^n$ consisting of all vectors that are annihilated by A, so

$$\ker A = \{\mathbf{z} \in \mathbb{R}^n \mid A\mathbf{z} = \mathbf{0}\} \subset \mathbb{R}^n.$$

The image is also known as the column space or the range of the matrix. By definition, a vector $\mathbf{b} \in \mathbb{R}^m$ belongs to $\operatorname{img} A$ if can be written as a linear combination,

$$\mathbf{b} = x_1\mathbf{v}_1 + \cdots + x_n\mathbf{v}_n$$

of the columns of $A = (\mathbf{v}_1\mathbf{v}_2 \ldots \mathbf{v}_n)$. The right-hand side of this equation equals the product $A\mathbf{x}$ of the matrix A with the column vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$, and, hence, $\mathbf{b} = A\mathbf{x}$ for some $\mathbf{x} \in \mathbb{R}^n$. Thus,

$$\operatorname{img} A = \{A\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n\} \subset \mathbb{R}^m,$$

and so a vector $\mathbf{b}$ lies in the image of A if and only if the linear system $A\mathbf{x} = \mathbf{b}$ has a solution.

A common alternative name for the kernel is the null space. The kernel or null space of A is the set of solutions $\mathbf{z}$ to the homogeneous system $A\mathbf{z} = \mathbf{0}$. The proof that $\ker A$ is a subspace requires us to verify the usual closure conditions: Suppose that $\mathbf{z}, \mathbf{w} \in \ker A$, so that $A\mathbf{z} = \mathbf{0} = A\mathbf{w}$. Then, by the compatibility of scalar and matrix multiplication, for any scalars c, d,

$$A(c\mathbf{z} + d\mathbf{w}) = cA\mathbf{z} + dA\mathbf{w} = \mathbf{0},$$

which implies that $c\mathbf{z} + d\mathbf{w} \in \ker A$.

3.4.1 Inner Product Spaces and Gram Matrices

A vector space equipped with an inner product and its associated norm, e.g., (3.1) and (3.2), is known as an *inner product space*. Other inner products are possible, and used. The general definition is:

Definition: An inner product on the real vector space V is a pairing that takes two vectors $\mathbf{v}, \mathbf{w} \in V$ and produces a real number $\langle \mathbf{v}, \mathbf{w} \rangle \in \mathbb{R}$. The inner product is required to satisfy the following three axioms for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$, and scalars $c, d \in \mathbb{R}$, termed, respectively, bilinearity, symmetry, and positivity:

(i) $\langle c\mathbf{u} + d\mathbf{v}, \mathbf{w} \rangle = c\langle \mathbf{u}, \mathbf{w} \rangle + d\langle \mathbf{v}, \mathbf{w} \rangle, \quad \langle \mathbf{u}, c\mathbf{v} + d\mathbf{w} \rangle = c\langle \mathbf{u}, \mathbf{v} \rangle + d\langle \mathbf{u}, \mathbf{w} \rangle.$
(ii) $\langle \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{w}, \mathbf{v} \rangle.$
(iii) $\langle \mathbf{v}, \mathbf{v} \rangle > 0$ whenever $\mathbf{v} \neq \mathbf{0}$, while $\langle \mathbf{0}, \mathbf{0} \rangle = 0.$

Every inner product gives rise to a norm that can be used to measure the magnitude or length of the elements of the underlying vector space. However, not every norm that is used

in analysis and applications arises from an inner product. To define a general norm on a vector space, we will extract those properties that do not directly rely on the inner product structure.

Definition: A norm on a vector space V assigns a non-negative real number $\|\mathbf{v}\|$ to each vector $\mathbf{v} \in V$, subject to the following axioms, valid for every $\mathbf{v}, \mathbf{w} \in V$ and $c \in \mathbb{R}$:

(i) Positivity: $\|\mathbf{v}\| \geq 0$, with $\|\mathbf{v}\| = 0$ if and only if $\mathbf{v} = \mathbf{0}$.
(ii) Homogeneity: $\|c\mathbf{v}\| = |c|\|\mathbf{v}\|$.
(iii) Triangle inequality: $\|\mathbf{v} + \mathbf{w}\| \leq \|\mathbf{v}\| + \|\mathbf{w}\|$.

Two important examples of norms that do not come from inner products (and that we will not use subsequently) are as follows. First, let $V = \mathbb{R}^n$. The 1-norm of a vector $\mathbf{v} = (v_1, v_2, \ldots, v_n)^T$ is defined as the sum of the absolute values of its entries:

$$\|\mathbf{v}\|_1 = |v_1| + |v_2| + \cdots + |v_n| .$$

The max- or ∞-norm is equal to its maximal entry (in absolute value):

$$\|\mathbf{v}\|_\infty = \max\{|v_1|, |v_2|, \ldots, |v_n|\} .$$

Every norm defines a distance between vector space elements, namely

$$d(\mathbf{v}, \mathbf{w}) = \|\mathbf{v} - \mathbf{w}\|. \tag{3.35}$$

For the standard dot product norm, we recover the usual notion of distance between points in Euclidean space. Other types of norms produce alternative (and sometimes quite useful) notions of distance that are, nevertheless, subject to all the familiar properties:

(a) Symmetry: $d(\mathbf{v}, \mathbf{w}) = d(\mathbf{w}, \mathbf{v})$;
(b) Positivity: $d(\mathbf{v}, \mathbf{w}) = 0$ if and only if $\mathbf{v} = \mathbf{w}$;
(c) Triangle inequality: $d(\mathbf{v}, \mathbf{w}) \leq d(\mathbf{v}, \mathbf{z}) + d(\mathbf{z}, \mathbf{w})$.

For the Euclidean norm, we wish to verify that the distance (3.35) obeys the triangle inequality, i.e., for $\mathbf{v}, \mathbf{w}, \mathbf{z} \in \mathbb{R}^n$, $\|\mathbf{v} - \mathbf{w}\| \leq \|\mathbf{v} - \mathbf{z}\| + \|\mathbf{z} - \mathbf{w}\|$. This is confirmed by recalling the triangle inequality (1.10) or (3.21), which states that, for $\mathbf{A}, \mathbf{B} \in \mathbb{R}^n$, $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$. Now apply this, taking $\mathbf{A} = \mathbf{v} - \mathbf{z}$ and $\mathbf{B} = \mathbf{z} - \mathbf{w}$.

Just as the distance between vectors measures how close they are to each other keeping in mind that this measure of proximity depends on the underlying choice of norm, so the distance between functions in a normed function space tells something about how close they are to each other, which is related, albeit subtly, to how close their graphs are. Thus, the norm serves to define the topology of the underlying vector space, which determines notions of open and closed sets, convergence, and so on.[25]

Suppose we are given an inner product $\langle \mathbf{x}, \mathbf{y} \rangle$ between vectors $\mathbf{x} = (x_1 x_2 \ldots x_n)^T$ and $\mathbf{y} = (y_1 y_2 \ldots y_n)^T$ in $\mathbb{R}^n$. Our goal is to determine its explicit formula. We begin by writing the vectors in terms of the standard basis vectors:

$$\mathbf{x} = x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n = \sum_{i=1}^{n} x_i\mathbf{e}_i, \quad \mathbf{y} = y_1\mathbf{e}_1 + \cdots + y_n\mathbf{e}_n = \sum_{j=1}^{n} y_j\mathbf{e}_j.$$

[25] For interested readers, there are several excellent books on (and with titles that include) Metric Space Theory. I highly recommend Heil's Metrics, Norms, Inner Products, and Operator Theory; Sasane's A Friendly Approach to Functional Analysis; and Lindstrøm's Spaces: An Introduction to Real Analysis.

To evaluate their inner product, we will appeal to the three basic axioms. We first employ bilinearity to expand

$$\langle \mathbf{x}, \mathbf{y} \rangle = \left\langle \sum_{i=1}^{n} x_i \mathbf{e}_i, \sum_{j=1}^{n} y_j \mathbf{e}_j \right\rangle = \sum_{i,j=1}^{n} x_i y_j \, \langle \mathbf{e}_i, \mathbf{e}_j \rangle .$$

Therefore,

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i,j=1}^{n} k_{ij} x_i y_j = \mathbf{x}^T K \mathbf{y}, \tag{3.36}$$

where K denotes the $n \times n$ matrix of inner products of the basis vectors, with entries

$$k_{ij} = \langle \mathbf{e}_i, \mathbf{e}_j \rangle , \quad i, j = 1, \ldots, n.$$

We conclude that any inner product must be expressed in the general bilinear form (3.36).

The two remaining inner product axioms will impose certain constraints on the inner product matrix K. Symmetry implies that

$$k_{ij} = \langle \mathbf{e}_i, \mathbf{e}_j \rangle = \langle \mathbf{e}_j, \mathbf{e}_i \rangle = k_{ji}, \quad i, j = 1, \ldots, n.$$

Consequently, the inner product matrix K must be symmetric, i.e., $K = K^T$. Conversely, symmetry of K ensures symmetry of the bilinear form:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T K \mathbf{y} = \big(\mathbf{x}^T K \mathbf{y} \big)^T = \mathbf{y}^T K^T \mathbf{x} = \mathbf{y}^T K \mathbf{x} = \langle \mathbf{y}, \mathbf{x} \rangle,$$

where the second equality follows from the fact that the quantity $\mathbf{x}^T K \mathbf{y}$ is a scalar, and hence equals its transpose. The final condition for an inner product is positivity:

$$\| \mathbf{x} \|^2 = \langle \mathbf{x}, \mathbf{x} \rangle = \mathbf{x}^T K \mathbf{x} = \sum_{i,j=1}^{n} k_{ij} x_i x_j \geq 0 \quad \text{for all} \quad \mathbf{x} \in \mathbb{R}^n,$$

with equality if and only if $\mathbf{x} = \mathbf{0}$. The precise meaning of this positivity condition on the matrix K is not so immediately evident, and so will be encapsulated in a definition.

Definition: An $n \times n$ matrix K is called *positive definite* if it is symmetric, and satisfies the positivity condition $\mathbf{x}^T K \mathbf{x} > 0$ for all $\mathbf{0} \neq \mathbf{x} \in \mathbb{R}^n$. We will sometimes write $K > 0$ to mean that K is a positive definite matrix.

Our preliminary analysis has resulted in the following general characterization of inner products on a finite-dimensional vector space:

Theorem: Every inner product on $\mathbb{R}^n$ is given by

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T K \mathbf{y} \quad \text{for} \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \tag{3.37}$$

where K is a symmetric, positive definite $n \times n$ matrix.

Definition: Given a symmetric matrix K, the homogeneous quadratic polynomial

$$q(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} = \sum_{i,j=1}^{n} k_{ij} x_i x_j \tag{3.38}$$

is known as a quadratic form on $\mathbb{R}^n$.

Definition: The quadratic form is called *positive definite* if $q(\mathbf{x}) > 0$ for all $\mathbf{0} \neq \mathbf{x} \in \mathbb{R}^n$.

So the quadratic form (3.38) is positive definite if and only if its coefficient matrix K is. It is easy to show that the coefficient matrix K in any quadratic form can be taken to be symmetric without any loss of generality. If a matrix is positive definite, then it is nonsingular.

A quadratic form and its associated symmetric coefficient matrix are called positive semi-definite if $q(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$, in which case we write $K \geq 0$.

Definition: Let V be an inner product space, and let $\mathbf{v}_1, \ldots, \mathbf{v}_n \in V$. The associated Gram matrix

$$K = \begin{pmatrix} \langle \mathbf{v}_1, \mathbf{v}_1 \rangle & \langle \mathbf{v}_1, \mathbf{v}_2 \rangle & \ldots & \langle \mathbf{v}_1, \mathbf{v}_n \rangle \\ \langle \mathbf{v}_2, \mathbf{v}_1 \rangle & \langle \mathbf{v}_2, \mathbf{v}_2 \rangle & \ldots & \langle \mathbf{v}_2, \mathbf{v}_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{v}_n, \mathbf{v}_1 \rangle & \langle \mathbf{v}_n, \mathbf{v}_2 \rangle & \ldots & \langle \mathbf{v}_n, \mathbf{v}_n \rangle \end{pmatrix} \tag{3.39}$$

is the $n \times n$ matrix whose entries are the inner products between the selected vector space elements.

Symmetry of the inner product implies symmetry of the Gram matrix:

$$k_{ij} = \langle \mathbf{v}_i, \mathbf{v}_j \rangle = \langle \mathbf{v}_j, \mathbf{v}_i \rangle = k_{ji}, \qquad \text{and hence} \qquad K^T = K. \tag{3.40}$$

In fact, the most direct method for producing positive definite and semi-definite matrices is through the Gram matrix construction.

Theorem: All Gram matrices are positive semi-definite. The Gram matrix above is positive definite if and only if $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are linearly independent.

Proof: To prove positive (semi-)definiteness of K, we need to examine the associated quadratic form $q(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} = \sum_{i,j=1}^n k_{ij} x_i x_j$. Substituting the values (3.40) for the matrix entries, we obtain

$$q(\mathbf{x}) = \sum_{i,j=1}^n \langle \mathbf{v}_i, \mathbf{v}_j \rangle \, x_i x_j.$$

Bilinearity of the inner product on V implies that we can assemble this summation into a single inner product

$$q(\mathbf{x}) = \left\langle \sum_{i=1}^n x_i \mathbf{v}_i, \sum_{j=1}^n x_j \mathbf{v}_j \right\rangle = \langle \mathbf{v}, \mathbf{v} \rangle = \|\mathbf{v}\|^2 \geq 0, \qquad \text{where} \qquad \mathbf{v} = \sum_{i=1}^n x_i \mathbf{v}_i$$

lies in the subspace of V spanned by the given vectors. This immediately proves that K is positive semi-definite. Moreover, $q(\mathbf{x}) = \|\mathbf{v}\|^2 > 0$ as long as $\mathbf{v} \neq \mathbf{0}$. If $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are linearly independent, then

$$\mathbf{v} = x_1 \mathbf{v}_1 + \cdots + x_n \mathbf{v}_n = \mathbf{0} \qquad \text{if and only if} \qquad x_1 = \cdots = x_n = 0,$$

and hence $q(\mathbf{x}) = 0$ if and only if $\mathbf{x} = \mathbf{0}$. This implies that $q(\mathbf{x})$ and hence K are positive definite.

In the case of the Euclidean dot product, the construction of the Gram matrix K can be directly implemented as follows. Given column vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{R}^m$, let us form the

$m \times n$ matrix $A = (\mathbf{v}_1 \mathbf{v}_2 \dots \mathbf{v}_n)$. In view of the identification between the dot product and multiplication of row and column vectors, the (i, j) entry of K is given as the product

$$k_{ij} = \mathbf{v}_i \cdot \mathbf{v}_j = \mathbf{v}_i^T \mathbf{v}_j$$

of the ith row of the transpose A^T and the jth column of A. In other words, the Gram matrix can be evaluated as a matrix product:

$$K = A^T A. \tag{3.41}$$

Changing the underlying inner product will, of course, change the Gram matrix. As noted in (3.37), every inner product on $\mathbb{R}^m$ has the form

$$\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T C \mathbf{w} \quad \text{for} \quad \mathbf{v}, \mathbf{w} \in \mathbb{R}^m, \tag{3.42}$$

where $C > 0$ is a symmetric, positive definite $m \times m$ matrix. Therefore, given n vectors $\mathbf{v}_1, \dots, \mathbf{v}_n \in \mathbb{R}^m$, the entries of the Gram matrix with respect to this inner product are

$$k_{ij} = \langle \mathbf{v}_i, \mathbf{v}_j \rangle = \mathbf{v}_i^T C \mathbf{v}_j.$$

If, as above, we assemble the column vectors into an $m \times n$ matrix $A = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_n \end{pmatrix}$, then the Gram matrix entry k_{ij} is obtained by multiplying the ith row of A^T by the jth column of the product matrix CA. Therefore, the Gram matrix based on the alternative inner product (3.42) is given by

$$K = A^T C A. \tag{3.43}$$

Recall the above theorem stating that all Gram matrices are positive semi-definite, and that a Gram matrix is positive definite if and only if $\mathbf{v}_1, \dots, \mathbf{v}_n$ are linearly independent. Thus, provided that the matrix A has full rank n, K is positive definite.

Theorem: Suppose A is an $m \times n$ matrix with linearly independent columns. Suppose C is any positive definite $m \times m$ matrix. Then the Gram matrix $K = A^T C A$ is a positive definite $n \times n$ matrix.

3.4.2 Orthogonal and Orthonormal Bases

Let V be a real inner product space. Recall that two elements $\mathbf{v}, \mathbf{w} \in V$ are called orthogonal if their inner product vanishes: $\langle \mathbf{v}, \mathbf{w} \rangle = 0$. In the case of vectors in Euclidean space, orthogonality under the dot product means that they meet at a right angle.

A particularly important configuration arises when V admits a basis consisting of mutually orthogonal elements.

Definition: A basis $\mathbf{u}_1, \dots, \mathbf{u}_n$ of an n-dimensional inner product space V is called orthogonal if $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0$ for all $i \neq j$. The basis is called orthonormal if, in addition, each vector has unit length: $\|\mathbf{u}_i\| = 1$, for all $i = 1, \dots, n$.

For the Euclidean space $\mathbb{R}^n$ equipped with the standard dot product, the simplest example of an orthonormal basis is the standard basis

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \quad \dots \quad \mathbf{e}_n = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

Orthogonality follows because $\mathbf{e}_i \cdot \mathbf{e}_j = 0$, for $i \neq j$, while $\|\mathbf{e}_i\| = 1$ implies normality. Since a basis cannot contain the zero vector, there is an easy way to convert an orthogonal basis to an orthonormal basis. Namely, we replace each basis vector with a unit vector pointing in the same direction.

Lemma: If $\mathbf{v}_1, \dots, \mathbf{v}_n$ is an orthogonal basis of a vector space V, then the normalized vectors $\mathbf{u}_i = \mathbf{v}_i / \|\mathbf{v}_i\|$, $i = 1, \dots, n$, form an orthonormal basis.

A useful observation is that every orthogonal collection of nonzero vectors is automatically linearly independent.

Proposition: Let $\mathbf{v}_1, \dots, \mathbf{v}_k \in V$ be nonzero, mutually orthogonal elements, so $\mathbf{v}_i \neq \mathbf{0}$ and $\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0$ for all $i \neq j$. Then $\mathbf{v}_1, \dots, \mathbf{v}_k$ are linearly independent.

Proof: Suppose

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

Let us take the inner product of this equation with any $\mathbf{v}_i$. Using linearity of the inner product and orthogonality, we compute

$$0 = \langle c_1 \mathbf{v}_1 + \cdots + c_k \mathbf{v}_k, \mathbf{v}_i \rangle = c_1 \langle \mathbf{v}_1, \mathbf{v}_i \rangle + \cdots + c_k \langle \mathbf{v}_k, \mathbf{v}_i \rangle = c_i \langle \mathbf{v}_i, \mathbf{v}_i \rangle = c_i \|\mathbf{v}_i\|^2 .$$

Therefore, given that $\mathbf{v}_i \neq \mathbf{0}$, we conclude that $c_i = 0$. Since this holds for all $i = 1, \dots, k$, the linear independence of $\mathbf{v}_1, \dots, \mathbf{v}_k$ follows.

As a direct corollary, we infer that every collection of nonzero orthogonal vectors forms a basis for its span:

Theorem: Suppose $\mathbf{v}_1, \dots, \mathbf{v}_n \in V$ are nonzero, mutually orthogonal elements of an inner product space V. Then $\mathbf{v}_1, \dots, \mathbf{v}_n$ form an orthogonal basis for their span $W = \operatorname{span}\{\mathbf{v}_1, \dots, \mathbf{v}_n\} \subset V$, which is therefore a subspace of dimension $n = \dim W$. In particular, if $\dim V = n$, then $\mathbf{v}_1, \dots, \mathbf{v}_n$ form a orthogonal basis for V.

What are the advantages of orthogonal and orthonormal bases? Once one has a basis of a vector space, a key issue is how to express other elements as linear combinations of the basis elements. That is, to find their coordinates in the prescribed basis. In general, this is not so easy, since it requires solving a system of linear equations. In high-dimensional situations arising in applications, computing the solution may require a considerable, if not infeasible, amount of time and effort.

However, if the basis is orthogonal, or, even better, orthonormal, then the change of basis computation requires almost no work. This is the crucial insight underlying the efficacy of both discrete and continuous Fourier analysis in signal, image, and video processing, least squares approximations, the statistical analysis of large data sets, and a multitude of other applications, both classical and modern.

Theorem: Let $\mathbf{u}_1, \dots, \mathbf{u}_n$ be an orthonormal basis for an inner product space V. Then one can write any element $\mathbf{v} \in V$ as a linear combination

$$\mathbf{v} = c_1 \mathbf{u}_1 + \cdots + c_n \mathbf{u}_n \tag{3.44}$$

in which its coordinates

$$c_i = \langle \mathbf{v}, \mathbf{u}_i \rangle , \quad i = 1, \dots, n, \tag{3.45}$$

are explicitly given as inner products. Moreover, its norm is given by the Pythagorean formula

$$\|\mathbf{v}\| = \sqrt{c_1^2 + \cdots + c_n^2} = \sqrt{\sum_{i=1}^{n} \langle \mathbf{v}, \mathbf{u}_i \rangle^2}, \tag{3.46}$$

namely, the square root of the sum of the squares of its orthonormal basis coordinates.

Proof: Let us compute the inner product of the element (3.44) with one of the basis vectors. Using the orthonormality conditions

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

and bilinearity of the inner product, we obtain

$$\langle \mathbf{v}, \mathbf{u}_i \rangle = \left\langle \sum_{j=1}^{n} c_j \mathbf{u}_j, \mathbf{u}_i \right\rangle = \sum_{j=1}^{n} c_j \langle \mathbf{u}_j, \mathbf{u}_i \rangle = c_i \|\mathbf{u}_i\|^2 = c_i.$$

To prove formula (3.46), we similarly expand

$$\|\mathbf{v}\|^2 = \langle \mathbf{v}, \mathbf{v} \rangle = \left\langle \sum_{j=1}^{n} c_i \mathbf{u}_i, \sum_{j=1}^{n} c_j \mathbf{u}_j \right\rangle = \sum_{i,j=1}^{n} c_i c_j \langle \mathbf{u}_i, \mathbf{u}_j \rangle = \sum_{i=1}^{n} c_i^2,$$

again making use of the orthonormality of the basis elements.

While passage from an orthogonal basis to its orthonormal version is elementary—one simply divides each basis element by its norm—we shall often find it more convenient to work directly with the unnormalized version. The next result provides the corresponding formula expressing a vector in terms of an orthogonal, but not necessarily orthonormal basis. The proof proceeds exactly as in the orthonormal case, and details are left to the reader.

Theorem: If $\mathbf{v}_1, \ldots, \mathbf{v}_n$ form an orthogonal basis, then the corresponding coordinates of a vector

$$\mathbf{v} = a_1 \mathbf{v}_1 + \cdots + a_n \mathbf{v}_n \quad \text{are given by} \quad a_i = \frac{\langle \mathbf{v}, \mathbf{v}_i \rangle}{\|\mathbf{v}_i\|^2}. \tag{3.47}$$

In this case, its norm can be computed using the formula

$$\|\mathbf{v}\|^2 = \sum_{i=1}^{n} a_i^2 \|\mathbf{v}_i\|^2 = \sum_{i=1}^{n} \left(\frac{\langle \mathbf{v}, \mathbf{v}_i \rangle}{\|\mathbf{v}_i\|} \right)^2.$$

3.4.3 Gram–Schmidt, Orthogonal Matrices, QR Factorization

Once we become convinced of the utility of orthogonal and orthonormal bases, a natural question arises: How can we construct them? A practical algorithm was first discovered by the French mathematician Pierre-Simon Laplace in the eighteenth century. Today the algorithm is known as the Gram-Schmidt process, after its rediscovery by Gram and the twentieth-century German mathematician Erhard Schmidt. The Gram-Schmidt process is one of the premier algorithms of applied and computational linear algebra.

Let W denote a finite-dimensional inner product space. (To begin with, you might wish to think of W as a subspace of $\mathbb{R}^m$, equipped with the standard Euclidean dot product, although the algorithm will be formulated in complete generality.) We assume that we already know some basis $\mathbf{w}_1, \dots, \mathbf{w}_n$ of W, where $n = \dim W$. Our goal is to use this information to construct an orthogonal basis $\mathbf{v}_1, \dots, \mathbf{v}_n$.

We will construct the orthogonal basis elements one by one. Since initially we are not worrying about normality, there are no conditions on the first orthogonal basis element $\mathbf{v}_1$, and so there is no harm in choosing $\mathbf{v}_1 = \mathbf{w}_1$. Note that $\mathbf{v}_1 \neq \mathbf{0}$, since $\mathbf{w}_1$ appears in the original basis. Starting with $\mathbf{w}_2$, the second basis vector $\mathbf{v}_2$ must be orthogonal to the first: $\langle \mathbf{v}_2, \mathbf{v}_1 \rangle = 0$. Let us try to arrange this by subtracting a suitable multiple of $\mathbf{v}_1$, and set

$$\mathbf{v}_2 = \mathbf{w}_2 - c\mathbf{v}_1,$$

where c is a scalar to be determined. The orthogonality condition

$$0 = \langle \mathbf{v}_2, \mathbf{v}_1 \rangle = \langle \mathbf{w}_2, \mathbf{v}_1 \rangle - c \langle \mathbf{v}_1, \mathbf{v}_1 \rangle = \langle \mathbf{w}_2, \mathbf{v}_1 \rangle - c \|\mathbf{v}_1\|^2$$

requires that $c = \langle \mathbf{w}_2, \mathbf{v}_1 \rangle / \|\mathbf{v}_1\|^2$, and therefore

$$\mathbf{v}_2 = \mathbf{w}_2 - \frac{\langle \mathbf{w}_2, \mathbf{v}_1 \rangle}{\|\mathbf{v}_1\|^2} \mathbf{v}_1.$$

Linear independence of $\mathbf{v}_1 = \mathbf{w}_1$ and $\mathbf{w}_2$ ensures that $\mathbf{v}_2 \neq \mathbf{0}$. Next, we construct

$$\mathbf{v}_3 = \mathbf{w}_3 - c_1\mathbf{v}_1 - c_2\mathbf{v}_2$$

by subtracting suitable multiples of the first two orthogonal basis elements from $\mathbf{w}_3$. We want $\mathbf{v}_3$ to be orthogonal to both $\mathbf{v}_1$ and $\mathbf{v}_2$. Since we already arranged that $\langle \mathbf{v}_1, \mathbf{v}_2 \rangle = 0$, this requires

$$0 = \langle \mathbf{v}_3, \mathbf{v}_1 \rangle = \langle \mathbf{w}_3, \mathbf{v}_1 \rangle - c_1 \langle \mathbf{v}_1, \mathbf{v}_1 \rangle, \quad 0 = \langle \mathbf{v}_3, \mathbf{v}_2 \rangle = \langle \mathbf{w}_3, \mathbf{v}_2 \rangle - c_2 \langle \mathbf{v}_2, \mathbf{v}_2 \rangle,$$

and hence

$$c_1 = \frac{\langle \mathbf{w}_3, \mathbf{v}_1 \rangle}{\|\mathbf{v}_1\|^2}, \quad c_2 = \frac{\langle \mathbf{w}_3, \mathbf{v}_2 \rangle}{\|\mathbf{v}_2\|^2}.$$

Therefore, the next orthogonal basis vector is given by the formula

$$\mathbf{v}_3 = \mathbf{w}_3 - \frac{\langle \mathbf{w}_3, \mathbf{v}_1 \rangle}{\|\mathbf{v}_1\|^2} \mathbf{v}_1 - \frac{\langle \mathbf{w}_3, \mathbf{v}_2 \rangle}{\|\mathbf{v}_2\|^2} \mathbf{v}_2.$$

Since $\mathbf{v}_1$ and $\mathbf{v}_2$ are linear combinations of $\mathbf{w}_1$ and $\mathbf{w}_2$, we must have $\mathbf{v}_3 \neq \mathbf{0}$, since otherwise, this would imply that $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3$ are linearly dependent, and hence could not come from a basis.

Continuing in the same manner, suppose we have already constructed the mutually orthogonal vectors $\mathbf{v}_1, \dots, \mathbf{v}_{k-1}$ as linear combinations of $\mathbf{w}_1, \dots, \mathbf{w}_{k-1}$. The next orthogonal basis element $\mathbf{v}_k$ will be obtained from $\mathbf{w}_k$ by subtracting off a suitable linear combination of the previous orthogonal basis elements:

$$\mathbf{v}_k = \mathbf{w}_k - c_1\mathbf{v}_1 - \cdots - c_{k-1}\mathbf{v}_{k-1}.$$

Since $\mathbf{v}_1, \dots, \mathbf{v}_{k-1}$ are already orthogonal, the orthogonality constraint

$$0 = \langle \mathbf{v}_k, \mathbf{v}_j \rangle = \langle \mathbf{w}_k, \mathbf{v}_j \rangle - c_j \langle \mathbf{v}_j, \mathbf{v}_j \rangle$$

requires

$$c_j = \frac{\langle \mathbf{w}_k, \mathbf{v}_j \rangle}{\|\mathbf{v}_j\|^2} \quad \text{for} \quad j = 1, \dots, k-1.$$

In this fashion, we establish the general Gram-Schmidt formula

$$\mathbf{v}_k = \mathbf{w}_k - \sum_{j=1}^{k-1} \frac{\langle \mathbf{w}_k, \mathbf{v}_j \rangle}{\|\mathbf{v}_j\|^2} \mathbf{v}_j, \quad k = 1, \dots, n. \tag{3.48}$$

The iterative Gram-Schmidt process (3.48), where we start with $\mathbf{v}_1 = \mathbf{w}_1$ and successively construct $\mathbf{v}_2, \dots, \mathbf{v}_n$, defines an explicit, recursive procedure for constructing the desired orthogonal basis vectors. If we are actually after an orthonormal basis $\mathbf{u}_1, \dots, \mathbf{u}_n$, we merely normalize the resulting orthogonal basis vectors, setting $\mathbf{u}_k = \mathbf{v}_k / \|\mathbf{v}_k\|$ for each $k = 1, \dots, n$.

Example 3.15 *The vectors*

$$\mathbf{w}_1 = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}, \quad \mathbf{w}_2 = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}, \quad \mathbf{w}_3 = \begin{pmatrix} 2 \\ -2 \\ 3 \end{pmatrix},$$

are readily seen to form a basis of $\mathbb{R}^3$. To construct an orthogonal basis (with respect to the standard dot product) using the Gram-Schmidt process, we begin by setting

$$\mathbf{v}_1 = \mathbf{w}_1 = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}$$

The next basis vector is

$$\mathbf{v}_2 = \mathbf{w}_2 - \frac{\mathbf{w}_2 \cdot \mathbf{v}_1}{\|\mathbf{v}_1\|^2} \mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} - \frac{-1}{3} \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} = \begin{pmatrix} \frac{4}{3} \\ \frac{1}{3} \\ \frac{5}{3} \end{pmatrix}.$$

The last orthogonal basis vector is

$$\mathbf{v}_3 = \mathbf{w}_3 - \frac{\mathbf{w}_3 \cdot \mathbf{v}_1}{\|\mathbf{v}_1\|^2} \mathbf{v}_1 - \frac{\mathbf{w}_3 \cdot \mathbf{v}_2}{\|\mathbf{v}_2\|^2} \mathbf{v}_2 = \begin{pmatrix} 2 \\ -2 \\ 3 \end{pmatrix} - \frac{-3}{3} \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} - \frac{7}{\frac{14}{3}} \begin{pmatrix} \frac{4}{3} \\ \frac{1}{3} \\ \frac{5}{3} \end{pmatrix} = \begin{pmatrix} 1 \\ -\frac{3}{2} \\ -\frac{1}{2} \end{pmatrix}.$$

The reader can easily validate the orthogonality of $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$. An orthonormal basis is obtained by dividing each vector by its length. Since

$$\|\mathbf{v}_1\| = \sqrt{3}, \quad \|\mathbf{v}_2\| = \sqrt{\frac{14}{3}}, \quad \|\mathbf{v}_3\| = \sqrt{\frac{7}{2}},$$

we produce the corresponding orthonormal basis vectors

$$\mathbf{u}_1 = \begin{pmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{3}} \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} \frac{4}{\sqrt{42}} \\ \frac{1}{\sqrt{42}} \\ \frac{5}{\sqrt{42}} \end{pmatrix}, \quad \mathbf{u}_3 = \begin{pmatrix} \frac{2}{\sqrt{14}} \\ -\frac{3}{\sqrt{14}} \\ -\frac{1}{\sqrt{14}} \end{pmatrix}. \qquad \blacksquare$$

Example 3.16 *Here is a typical problem: find an orthonormal basis, with respect to the dot product, for the subspace $W \subset \mathbb{R}^4$ consisting of all vectors that are orthogonal to the given vector $\mathbf{a} = (1, 2, -1, -3)^T$. The first task is to find a basis for the subspace. Now, a vector $\mathbf{x} = (x_1, x_2, x_3, x_4)^T$ is orthogonal to $\mathbf{a}$ if and only if*

$$\mathbf{x} \cdot \mathbf{a} = x_1 + 2x_2 - x_3 - 3x_4 = 0.$$

Solving this homogeneous linear system by the usual method, we observe that the free variables are x_2, x_3, x_4, and so a (non-orthogonal) basis for the subspace is

$$\mathbf{w}_1 = \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{w}_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{w}_3 = \begin{pmatrix} 3 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

To obtain an orthogonal basis, we apply the Gram-Schmidt process. First,

$$\mathbf{v}_1 = \mathbf{w}_1 = \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

The next element is

$$\mathbf{v}_2 = \mathbf{w}_2 - \frac{\mathbf{w}_2 \cdot \mathbf{v}_1}{\|\mathbf{v}_1\|^2}\mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} - \frac{-2}{5}\begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{5} \\ \frac{2}{5} \\ 1 \\ 0 \end{pmatrix}.$$

The last element of our orthogonal basis is

$$\mathbf{v}_3 = \mathbf{w}_3 - \frac{\mathbf{w}_3 \cdot \mathbf{v}_1}{\|\mathbf{v}_1\|^2}\mathbf{v}_1 - \frac{\mathbf{w}_3 \cdot \mathbf{v}_2}{\|\mathbf{v}_2\|^2}\mathbf{v}_2 = \begin{pmatrix} 3 \\ 0 \\ 0 \\ 1 \end{pmatrix} - \frac{-6}{5}\begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix} - \frac{\frac{3}{5}}{\frac{6}{5}}\begin{pmatrix} \frac{1}{5} \\ \frac{2}{5} \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ 1 \\ -\frac{1}{2} \\ 1 \end{pmatrix}.$$

An orthonormal basis can then be obtained by dividing each $\mathbf{v}_i$ by its length:

$$\mathbf{u}_1 = \begin{pmatrix} -\frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} \frac{1}{\sqrt{30}} \\ \frac{2}{\sqrt{30}} \\ \frac{5}{\sqrt{30}} \\ 0 \end{pmatrix}, \quad \mathbf{u}_3 = \begin{pmatrix} \frac{1}{\sqrt{10}} \\ \frac{2}{\sqrt{10}} \\ -\frac{1}{\sqrt{10}} \\ \frac{2}{\sqrt{10}} \end{pmatrix}. \qquad \blacksquare$$

With the basic Gram-Schmidt algorithm now in hand, it is worth looking at a couple of reformulations that have both practical and theoretical advantages. The first can be used to construct the orthonormal basis vectors $\mathbf{u}_1, \dots, \mathbf{u}_n$ directly from the basis $\mathbf{w}_1, \dots, \mathbf{w}_n$.

We begin by replacing each orthogonal basis vector in the basic Gram-Schmidt formula (3.48) by its normalized version $\mathbf{u}_j = \mathbf{v}_j / \|\mathbf{v}_j\|$. The original basis vectors can be expressed in terms of the orthonormal basis via a "triangular" system

$$\begin{aligned} \mathbf{w}_1 &= r_{11}\mathbf{u}_1 \\ \mathbf{w}_2 &= r_{12}\mathbf{u}_1 + r_{22}\mathbf{u}_2, \\ \mathbf{w}_3 &= r_{13}\mathbf{u}_1 + r_{23}\mathbf{u}_2 + r_{33}\mathbf{u}_3, \\ &\vdots \quad \vdots \quad \vdots \quad \ddots \\ \mathbf{w}_n &= r_{1n}\mathbf{u}_1 + r_{2n}\mathbf{u}_2 + \cdots + r_{nn}\mathbf{u}_n. \end{aligned} \tag{3.49}$$

Before proceeding, it is useful to cast this in matrix terms. It will be used below by the authors in (3.55). For illustration, take $n = 2$, so that

$$\begin{bmatrix} w_{1,1} \\ w_{2,1} \end{bmatrix} = \mathbf{w}_1 = r_{11}\mathbf{u}_1 = \begin{bmatrix} r_{11}u_{1,1} \\ r_{11}u_{2,1} \end{bmatrix}, \quad \begin{bmatrix} w_{1,2} \\ w_{2,2} \end{bmatrix} = \mathbf{w}_2 = r_{12}\mathbf{u}_1 + r_{22}\mathbf{u}_2 = \begin{bmatrix} r_{12}u_{1,1} + r_{22}u_{1,2} \\ r_{12}u_{2,1} + r_{22}u_{2,2} \end{bmatrix}$$

or, with $\mathbf{W} = [\mathbf{w}_1\ \mathbf{w}_2]$ and $\mathbf{U} = [\mathbf{u}_1\ \mathbf{u}_2]$,

$$\mathbf{W} = \begin{bmatrix} w_{1,1} & w_{1,2} \\ w_{2,1} & w_{2,2} \end{bmatrix} = \begin{bmatrix} r_{11}u_{1,1} & r_{12}u_{1,1} + r_{22}u_{1,2} \\ r_{11}u_{2,1} & r_{12}u_{2,1} + r_{22}u_{2,2} \end{bmatrix} = \begin{bmatrix} u_{1,1} & u_{1,2} \\ u_{2,1} & u_{2,2} \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{bmatrix} = \mathbf{UR}, \tag{3.50}$$

where $\mathbf{R}$ is the 2×2 indicated matrix, and the case of general n is clear.

The coefficients r_{ij} can, in fact, be computed directly from these formulas. Indeed, taking the inner product of the equation for $\mathbf{w}_j$ with the orthonormal basis vector $\mathbf{u}_i$ for $i \le j$, we obtain, in view of the orthonormality constraints,

$$\langle \mathbf{w}_j, \mathbf{u}_i \rangle = \langle r_{1j}\mathbf{u}_1 + \cdots + r_{jj}\mathbf{u}_j, \mathbf{u}_i \rangle = r_{1j} \langle \mathbf{u}_1, \mathbf{u}_i \rangle + \cdots + r_{jj} \langle \mathbf{u}_j, \mathbf{u}_i \rangle = r_{ij},$$

and hence

$$r_{ij} = \langle \mathbf{w}_j, \mathbf{u}_i \rangle . \tag{3.51}$$

On the other hand,

$$\|\mathbf{w}_j\|^2 = \|r_{1j}\mathbf{u}_1 + \cdots + r_{jj}\mathbf{u}_j\|^2 = r_{1j}^2 + \cdots + r_{j-1,j}^2 + r_{jj}^2. \tag{3.52}$$

The pair of equations (3.51) and (3.52) can be rearranged to devise a recursive procedure to compute the orthonormal basis. We begin by setting $r_{11} = \|\mathbf{w}_1\|$ and so $\mathbf{u}_1 = \mathbf{w}_1/r_{11}$. At each subsequent stage $j \ge 2$, we assume that we have already constructed $\mathbf{u}_1, \dots, \mathbf{u}_{j-1}$. We then compute

$$r_{ij} = \langle \mathbf{w}_j, \mathbf{u}_i \rangle , \qquad \text{for each} \qquad i = 1, \dots, j-1. \tag{3.53}$$

We obtain the next orthonormal basis vector $\mathbf{u}_j$ by computing

$$r_{jj} = \sqrt{\|\mathbf{w}_j\|^2 - r_{1j}^2 - \cdots - r_{j-1,j}^2}, \quad \mathbf{u}_j = \frac{\mathbf{w}_j - r_{1j}\mathbf{u}_1 - \cdots - r_{j-1,j}\mathbf{u}_{j-1}}{r_{jj}}. \tag{3.54}$$

Running through the formulas (3.53) and (3.54) for $j = 1, \dots, n$ leads to the same orthonormal basis $\mathbf{u}_1, \dots, \mathbf{u}_n$ produced by the previous version of the Gram-Schmidt procedure.

In practical, large-scale computations, both versions of the Gram-Schmidt process suffer from a serious flaw. They are subject to numerical instabilities, and so accumulating round-off errors may seriously corrupt the computations, leading to inaccurate, non-orthogonal vectors. Fortunately, there is a simple rearrangement of the calculation that ameliorates this difficulty and leads to the numerically robust algorithm that is most often used in practice. The idea is to treat the vectors simultaneously rather than sequentially, making full use of the orthonormal basis vectors as they arise. More specifically, the algorithm begins as before—we take $\mathbf{u}_1 = \mathbf{w}_1 / \|\mathbf{w}_1\|$. We then subtract off the appropriate multiples of $\mathbf{u}_1$ from all of the remaining basis vectors so as to arrange their orthogonality to $\mathbf{u}_1$. This is accomplished by setting

$$\mathbf{w}_k^{(2)} = \mathbf{w}_k - \langle \mathbf{w}_k, \mathbf{u}_1 \rangle \mathbf{u}_1 \qquad \text{for} \qquad k = 2, \dots, n.$$

The second orthonormal basis vector $\mathbf{u}_2 = \mathbf{w}_2^{(2)} / \left\| \mathbf{w}_2^{(2)} \right\|$ is then obtained by normalizing. We next modify the remaining $\mathbf{w}_3^{(2)}, \dots, \mathbf{w}_n^{(2)}$ to produce vectors

$$\mathbf{w}_k^{(3)} = \mathbf{w}_k^{(2)} - \left\langle \mathbf{w}_k^{(2)}, \mathbf{u}_2 \right\rangle \mathbf{u}_2, \quad k = 3, \dots, n,$$

that are orthogonal to both $\mathbf{u}_1$ and $\mathbf{u}_2$. Then $\mathbf{u}_3 = \mathbf{w}_3^{(3)} / \left\| \mathbf{w}_3^{(3)} \right\|$ is the next orthonormal basis element, and the process continues. The full algorithm starts with the initial basis vectors $\mathbf{w}_j = \mathbf{w}_j^{(1)}$, $j = 1, \dots, n$, and then recursively computes, for $j = 1, \dots, n$ and $k = j+1, \dots, n$,

$$\mathbf{u}_j = \frac{\mathbf{w}_j^{(j)}}{\left\| \mathbf{w}_j^{(j)} \right\|}, \quad \mathbf{w}_k^{(j+1)} = \mathbf{w}_k^{(j)} - \left\langle \mathbf{w}_k^{(j)}, \mathbf{u}_j \right\rangle \mathbf{u}_j.$$

(In the final phase, when $j = n$, the second formula is no longer needed.) The result is a numerically stable computation of the same orthonormal basis vectors $\mathbf{u}_1, \dots, \mathbf{u}_n$.

Matrices whose columns form an orthonormal basis of $\mathbb{R}^n$ relative to the standard Euclidean dot product play a distinguished role. Such "orthogonal matrices" appear in a wide range of applications in geometry, physics, quantum mechanics, crystallography, partial differential equations, symmetry theory, and special functions. Rotational motions of bodies in three-dimensional space are described by orthogonal matrices, and hence they lie at the foundations of rigid body mechanics, including satellites, airplanes, drones, and underwater vehicles, as well as three-dimensional computer graphics and animation for video games and movies. Furthermore, orthogonal matrices are an essential ingredient in one of the most important methods of numerical linear algebra: the QR algorithm for computing eigenvalues of matrices.

Definition: A square matrix Q is called orthogonal if it satisfies $Q^T Q = Q Q^T = \mathrm{I}$.

This is referred to as the orthogonality condition, or requirement. It implies that one can easily invert an orthogonal matrix: $Q^{-1} = Q^T$. In fact, the two conditions are equivalent, and hence a matrix is orthogonal if and only if its inverse is equal to its transpose. In particular, the identity matrix I is orthogonal. Also note that, if Q is orthogonal, then so is Q^T. The second important characterization of orthogonal matrices relates them directly to orthonormal bases.

Proposition: A matrix Q is orthogonal if and only if its columns form an orthonormal basis with respect to the Euclidean dot product on $\mathbb{R}^n$.

> Proof: Let $\mathbf{u}_1, \dots, \mathbf{u}_n$ be the columns of Q. Then $\mathbf{u}_1^T, \dots, \mathbf{u}_n^T$ are the rows of the transposed matrix Q^T. The (i, j) entry of the product $Q^T Q$ is given as the product of the ith row of Q^T and the jth column of Q. Thus, the orthogonality requirement implies
>
> $$\mathbf{u}_i \cdot \mathbf{u}_j = \mathbf{u}_i^T \mathbf{u}_j = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases}$$
>
> which are precisely the conditions for $\mathbf{u}_1, \dots, \mathbf{u}_n$ to form an orthonormal basis.

In particular, the columns of the identity matrix produce the standard basis $\mathbf{e}_1, \dots, \mathbf{e}_n$ of $\mathbb{R}^n$. Also, the rows of an orthogonal matrix Q also produce an (in general different) orthonormal basis.

Warning: Technically, we should be referring to an "orthonormal" matrix, not an "orthogonal" matrix. But the terminology is so standard throughout mathematics and physics that we have no choice but to adopt it here. There is no commonly accepted name for a matrix whose columns form an orthogonal but not orthonormal basis.

Example 3.17 *A* 2×2 *matrix* $Q = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ *is orthogonal if and only if its columns* $\mathbf{u}_1 = \begin{pmatrix} a \\ c \end{pmatrix}, \mathbf{u}_2 = \begin{pmatrix} b \\ d \end{pmatrix}$, *form an orthonormal basis of* $\mathbb{R}^2$. *Equivalently, the requirement*

$$Q^T Q = \begin{pmatrix} a & c \\ b & d \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a^2 + c^2 & ab + cd \\ ab + cd & b^2 + d^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

implies that its entries must satisfy the algebraic equations $a^2+c^2 = 1$, $ab+cd = 0$, $b^2+d^2 = 1$. *The first and last equations say that the points* $(a, c)^T$ *and* $(b, d)^T$ *lie on the unit circle in* $\mathbb{R}^2$, *and so*

$$a = \cos\theta, \quad c = \sin\theta, \quad b = \cos\psi, \quad d = \sin\psi$$

for some choice of angles θ, ψ. *The remaining orthogonality condition is*

$$0 = ab + cd = \cos\theta\cos\psi + \sin\theta\sin\psi = \cos(\theta - \psi),$$

which implies that θ *and* ψ *differ by a right angle:* $\psi = \theta \pm \frac{1}{2}\pi$. *The* $\pm$ *sign leads to two cases: It is useful to recall the relations in (3.24).*

$$b = -\sin\theta, \quad d = \cos\theta, \qquad \text{or} \qquad b = \sin\theta, \quad d = -\cos\theta.$$

As a result, every 2×2 *orthogonal matrix has one of two possible forms*

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}, \qquad \text{where } 0 \le \theta < 2\pi.$$

The corresponding orthonormal bases are illustrated in Figure 28. ■

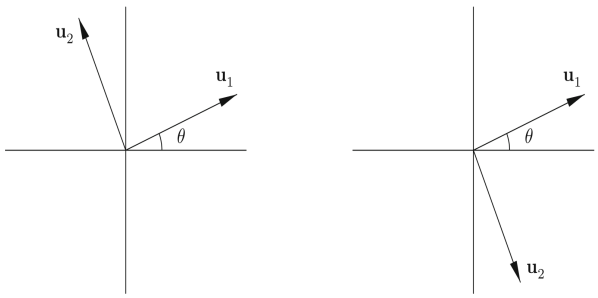


Figure 28: Orthonormal bases in $\mathbb{R}^2$

Lemma: An orthogonal matrix Q has determinant $\det Q = \pm 1$.

Proof: Taking the determinant of $Q^T Q = Q Q^T = I$ shows that

$$1 = \det I = \det\left(Q^T Q\right) = \det Q^T \det Q = (\det Q)^2,$$

which immediately proves the lemma.

An orthogonal matrix is called proper or special if it has determinant $+1$. An improper orthogonal matrix, with determinant -1, corresponds to a left handed basis that lives in a mirror-image world.

Proposition: The product of two orthogonal matrices is also orthogonal.

Proof: If

$$Q_1^T Q_1 = I = Q_2^T Q_2, \quad \text{then} \quad (Q_1 Q_2)^T (Q_1 Q_2) = Q_2^T Q_1^T Q_1 Q_2 = Q_2^T Q_2 = I,$$

and so the product matrix $Q_1 Q_2$ is also orthogonal.

This multiplicative property, combined with the fact that the inverse of an orthogonal matrix is also orthogonal, says that the set of all orthogonal matrices forms a group. The orthogonal group lies at the foundation of everyday Euclidean geometry, as well as rigid body mechanics, atomic structure and chemistry, computer graphics and animation, and many other areas.

The Gram-Schmidt procedure for orthonormalizing bases of $\mathbb{R}^n$ can be reinterpreted as a matrix factorization. This is more subtle than the LU factorization that resulted from Gaussian Elimination, but is of comparable significance, and is used in a broad range of applications in mathematics, statistics, physics, engineering, and numerical analysis.

Let $\mathbf{w}_1, \dots, \mathbf{w}_n$ be a basis of $\mathbb{R}^n$, and let $\mathbf{u}_1, \dots, \mathbf{u}_n$ be the corresponding orthonormal basis that results from any one of the three implementations of the Gram-Schmidt process. We assemble both sets of column vectors to form nonsingular $n \times n$ matrices

$$A = \begin{pmatrix} \mathbf{w}_1 & \mathbf{w}_2 & \dots & \mathbf{w}_n \end{pmatrix}, \quad Q = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_n \end{pmatrix}.$$

Since the $\mathbf{u}_i$ form an orthonormal basis, Q is an orthogonal matrix. The Gram-Schmidt equations (3.49) can be recast into an equivalent matrix form: Recall (3.50)

$$A = QR, \quad \text{where} \quad R = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ 0 & r_{22} & \dots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & r_{nn} \end{pmatrix} \tag{3.55}$$

is an upper triangular matrix whose entries are the coefficients in (3.53) and (3.54). Since the Gram-Schmidt process works on any basis, the only requirement on the matrix A is that its columns form a basis of $\mathbb{R}^n$, and hence A can be any nonsingular matrix. We have therefore established the celebrated QR factorization of nonsingular matrices.

Theorem: Every nonsingular matrix can be factored, $A = QR$, into the product of an orthogonal matrix Q and an upper triangular matrix R. The factorization is unique if R is positive upper triangular, meaning that all its diagonal entries of are positive.

3.4.4 Orthogonal Projections and Orthogonal Subspaces

Definition: A vector $\mathbf{z} \in V$ is said to be orthogonal to the subspace $W \subset V$ if it is orthogonal to every vector in W, so $\langle \mathbf{z}, \mathbf{w} \rangle = 0$ for all $\mathbf{w} \in W$.

Given a basis $\mathbf{w}_1, \dots, \mathbf{w}_n$ of the subspace W, we note that $\mathbf{z}$ is orthogonal to W if and only if it is orthogonal to every basis vector: $\langle \mathbf{z}, \mathbf{w}_i \rangle = 0$ for $i = 1, \dots, n$. Indeed, any

other vector in W has the form $\mathbf{w} = c_1\mathbf{w}_1 + \cdots + c_n\mathbf{w}_n$, and hence, by linearity, $\langle \mathbf{z}, \mathbf{w} \rangle = c_1 \langle \mathbf{z}, \mathbf{w}_1 \rangle + \cdots + c_n \langle \mathbf{z}, \mathbf{w}_n \rangle = 0$, as required.

Definition: The orthogonal projection of $\mathbf{v}$ onto the subspace W is the element $\mathbf{w} \in W$ that makes the difference $\mathbf{z} = \mathbf{v} - \mathbf{w}$ orthogonal to W.

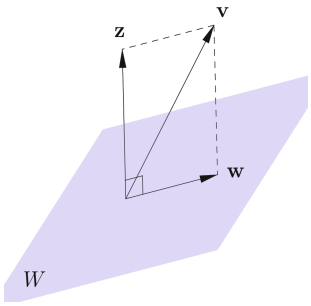


Figure 29: The orthogonal projection of a vector onto a subspace

The geometric configuration underlying orthogonal projection is sketched in Figure 29. As we shall see, the orthogonal projection is unique. Note that $\mathbf{v} = \mathbf{w} + \mathbf{z}$ is the sum of its orthogonal projection $\mathbf{w} \in V$ and the perpendicular vector $\mathbf{z} \perp W$. The explicit construction is greatly simplified by taking an orthonormal basis of the subspace, which, if necessary, can be arranged by applying the Gram-Schmidt process to a known basis.

Theorem: Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be an orthonormal basis for the subspace $W \subset V$. Then the orthogonal projection of $\mathbf{v} \in V$ onto $\mathbf{w} \in W$ is given by

$$\mathbf{w} = c_1\mathbf{u}_1 + \cdots + c_n\mathbf{u}_n \qquad \text{where} \qquad c_i = \langle \mathbf{v}, \mathbf{u}_i \rangle, \quad i = 1, \ldots, n. \tag{3.56}$$

Proof: First, since $\mathbf{u}_1, \ldots, \mathbf{u}_n$ form a basis of the subspace, the orthogonal projection element must be some linear combination thereof: $\mathbf{w} = c_1\mathbf{u}_1 + \cdots + c_n\mathbf{u}_n$. The definition of the orthogonal projection of $\mathbf{v}$ onto the subspace W requires that the difference $\mathbf{z} = \mathbf{v} - \mathbf{w}$ be orthogonal to W, and, as noted above, it suffices to check orthogonality to the basis vectors. By our orthonormality assumption,

$$\begin{aligned} 0 = \langle \mathbf{z}, \mathbf{u}_i \rangle &= \langle \mathbf{v} - \mathbf{w}, \mathbf{u}_i \rangle = \langle \mathbf{v} - c_1\mathbf{u}_1 - \cdots - c_n\mathbf{u}_n, \mathbf{u}_i \rangle \\ &= \langle \mathbf{v}, \mathbf{u}_i \rangle - c_1 \langle \mathbf{u}_1, \mathbf{u}_i \rangle - \cdots - c_n \langle \mathbf{u}_n, \mathbf{u}_i \rangle = \langle \mathbf{v}, \mathbf{u}_i \rangle - c_i. \end{aligned}$$

The coefficients $c_i = \langle \mathbf{v}, \mathbf{u}_i \rangle$ of the orthogonal projection $\mathbf{w}$ are thus uniquely prescribed by the orthogonality requirement, which thereby proves its uniqueness.

More generally, if we employ an orthogonal basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ for the subspace W, then the same argument demonstrates that the orthogonal projection of $\mathbf{v}$ onto W is given by

$$\mathbf{w} = a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n, \qquad \text{where} \qquad a_i = \frac{\langle \mathbf{v}, \mathbf{v}_i \rangle}{\|\mathbf{v}_i\|^2}, \quad i = 1, \ldots, n. \tag{3.57}$$

We could equally well replace the orthogonal basis by the orthonormal basis obtained by dividing each vector by its length: $\mathbf{u}_i = \mathbf{v}_i / \|\mathbf{v}_i\|$. The reader should be able to prove that the two formulas (3.56) and (3.57) for the orthogonal projection yield the same vector $\mathbf{w}$.

An intriguing observation is that the coefficients in the orthogonal projection formulas (3.56) and (3.57) coincide with the formulas (3.45) and (3.47) for writing a vector in terms of

an orthonormal or orthogonal basis. Indeed, if $\mathbf{v}$ were an element of W, then it would coincide with its orthogonal projection, $\mathbf{w} = \mathbf{v}$. As a result, the orthogonal projection formula include the orthogonal basis formula as a special case.

It is also worth noting that the same formulae occur in the Gram-Schmidt algorithm, cf. (3.48). This observation leads to a useful geometric interpretation of the Gram-Schmidt construction. For each $k = 1, \dots, n$, let

$$W_k = \operatorname{span}\{\mathbf{w}_1, \dots, \mathbf{w}_k\} = \operatorname{span}\{\mathbf{v}_1, \dots, \mathbf{v}_k\} = \operatorname{span}\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$$

denote the k-dimensional subspace spanned by the first k basis elements, which is the same as that spanned by their orthogonalized and orthonormalized counterparts. In view of (4.41), the basic Gram-Schmidt formula (3.48) can be re-expressed in the form $\mathbf{v}_k = \mathbf{w}_k - \mathbf{p}_k$, where $\mathbf{p}_k$ is the orthogonal projection of $\mathbf{w}_k$ onto the subspace W_{k-1}. The resulting vector $\mathbf{v}_k$ is, by construction, orthogonal to the subspace, and hence orthogonal to all of the previous basis elements, which serves to rejustify the Gram-Schmidt construction.

We now extend the notion of orthogonality from individual elements to entire subspaces of an inner product space V.

Definition: Two subspaces $W, Z \subset V$ are called orthogonal if every vector in W is orthogonal to every vector in Z.

In other words, W and Z are orthogonal subspaces if and only if $\langle \mathbf{w}, \mathbf{z} \rangle = 0$ for every $\mathbf{w} \in W$ and $\mathbf{z} \in Z$. In practice, one only needs to check orthogonality of basis elements, or, more generally, spanning sets.

Lemma: If $\mathbf{w}_1, \dots, \mathbf{w}_k$ span W and $\mathbf{z}_1, \dots, \mathbf{z}_l$ span Z, then W and Z are orthogonal subspaces if and only if $\langle \mathbf{w}_i, \mathbf{z}_j \rangle = 0$ for all $i = 1, \dots, k$ and $j = 1, \dots, l$. The proof of this lemma is left to the reader.

Example 3.18 *Let $V = \mathbb{R}^3$ have the ordinary dot product. Then the plane $W \subset \mathbb{R}^3$ defined by the equation $2x - y + 3z = 0$ is orthogonal to the line Z spanned by its normal vector $\mathbf{n} = (2, -1, 3)^T$. Indeed, every $\mathbf{w} = (x, y, z)^T \in W$ satisfies the orthogonality condition $\mathbf{w} \cdot \mathbf{n} = 2x - y + 3z = 0$, which is simply the equation for the plane.* ■

Example 3.19 *Let W be the span of $\mathbf{w}_1 = (1, -2, 0, 1)^T, \mathbf{w}_2 = (3, -5, 2, 1)^T$, and let Z be the span of the vectors $\mathbf{z}_1 = (3, 2, 0, 1)^T, \mathbf{z}_2 = (1, 0, -1, -1)^T$. We find that $\mathbf{w}_1 \cdot \mathbf{z}_1 = \mathbf{w}_1 \cdot \mathbf{z}_2 = \mathbf{w}_2 \cdot \mathbf{z}_1 = \mathbf{w}_2 \cdot \mathbf{z}_2 = 0$, and so W and Z are orthogonal two-dimensional subspaces of $\mathbb{R}^4$ under the Euclidean dot product.* ■

Definition: The orthogonal complement of a subspace $W \subset V$, denoted $W^\perp$, is defined as the set of all vectors that are orthogonal to W:

$$W^\perp = \{\mathbf{v} \in V \mid \langle \mathbf{v}, \mathbf{w} \rangle = 0 \text{ for all } \mathbf{w} \in W\}.$$

If W is the one-dimensional subspace (line) spanned by a single vector $\mathbf{w} \neq \mathbf{0}$, then we also denote $W^\perp$ by $\mathbf{w}^\perp$. One easily checks that the orthogonal complement $W^\perp$ is also a subspace. Moreover, $W \cap W^\perp = \{\mathbf{0}\}$. Keep in mind that the orthogonal complement will depend upon which inner product is being used.

Example 3.20 *Let* $W = \{(t, 2t, 3t)^T \mid t \in \mathbb{R}\}$ *be the line (one-dimensional subspace) in the direction of the vector* $\mathbf{w}_1 = (1, 2, 3)^T \in \mathbb{R}^3$. *Under the dot product, its orthogonal complement* $W^\perp = \mathbf{w}_1^\perp$ *is the plane passing through the origin having normal vector* $\mathbf{w}_1$. *In other words,* $\mathbf{z} = (x, y, z)^T \in W^\perp$ *if and only if*

$$\mathbf{z} \cdot \mathbf{w}_1 = x + 2y + 3z = 0.$$

Thus, $W^\perp$ *is characterized as the solution space of this previous homogeneous linear equation, or, equivalently, the kernel of the* 1×3 *matrix* $A = \mathbf{w}_1^T = \begin{pmatrix} 1 & 2 & 3 \end{pmatrix}$. *We can write the general solution in the form*

$$\mathbf{z} = \begin{pmatrix} -2y - 3z \\ y \\ z \end{pmatrix} = y \begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix} + z \begin{pmatrix} -3 \\ 0 \\ 1 \end{pmatrix} = y\mathbf{z}_1 + z\mathbf{z}_2,$$

where y, z *are the free variables. The indicated vectors* $\mathbf{z}_1 = (-2, 1, 0)^T, \mathbf{z}_2 = (-3, 0, 1)^T$, *form a (non-orthogonal) basis for the orthogonal complement* $W^\perp$. ■

Proposition: Suppose that $W \subset V$ is a finite-dimensional subspace of an inner product space. Then every vector $\mathbf{v} \in V$ can be uniquely decomposed into $\mathbf{v} = \mathbf{w} + \mathbf{z}$, where $\mathbf{w} \in W$ and $\mathbf{z} \in W^\perp$.

> Proof: We let $\mathbf{w} \in W$ be the orthogonal projection of $\mathbf{v}$ onto W. Then $\mathbf{z} = \mathbf{v} - \mathbf{w}$ is, by definition, orthogonal to W and hence belongs to $W^\perp$. Note that $\mathbf{z}$ can be viewed as the orthogonal projection of $\mathbf{v}$ onto the complementary subspace $W^\perp$ (provided it is finite-dimensional). If we are given two such decompositions, $\mathbf{v} = \mathbf{w} + \mathbf{z} = \widetilde{\mathbf{w}} + \widetilde{\mathbf{z}}$, then $\mathbf{w} - \widetilde{\mathbf{w}} = \widetilde{\mathbf{z}} - \mathbf{z}$. The left-hand side of this equation lies in W, while the right-hand side belongs to $W^\perp$. But, as we already noted, the only vector that belongs to both W and $W^\perp$ is the zero vector. Thus, $\mathbf{w} - \widetilde{\mathbf{w}} = \mathbf{0} = \widetilde{\mathbf{z}} - \mathbf{z}$, so $\mathbf{w} = \widetilde{\mathbf{w}}$ and $\mathbf{z} = \widetilde{\mathbf{z}}$, which proves uniqueness.

As a consequence of this previous proposition, we have:

Proposition: If W is a finite-dimensional subspace of an inner product space, then $\left(W^\perp\right)^\perp = W$.

In a finite-dimensional inner product space, a subspace and its orthogonal complement have complementary dimensions:

Proposition: If $W \subset V$ is a subspace with $\dim W = n$ and $\dim V = m$, then $\dim W^\perp = m - n$.

Example 3.21 *Let* $W \subset \mathbb{R}^4$ *be the two-dimensional subspace spanned by the orthogonal vectors* $\mathbf{w}_1 = (1, 1, 0, 1)^T$ *and* $\mathbf{w}_2 = (1, 1, 1, -2)^T$. *Its orthogonal complement* $W^\perp$ *(with respect to the Euclidean dot product) is the set of all vectors* $\mathbf{v} = (x, y, z, w)^T$ *that satisfy the linear system*

$$\mathbf{v} \cdot \mathbf{w}_1 = x + y + w = 0, \quad \mathbf{v} \cdot \mathbf{w}_2 = x + y + z - 2w = 0.$$

Applying the usual algorithm (the free variables are y *and* w*), we find that the solution space is spanned by*

$$\mathbf{z}_1 = (-1, 1, 0, 0)^T, \quad \mathbf{z}_2 = (-1, 0, 3, 1)^T,$$

which form a non-orthogonal basis for $W^\perp$. An orthogonal basis

$$\mathbf{y}_1 = \mathbf{z}_1 = (-1, 1, 0, 0)^T, \quad \mathbf{y}_2 = \mathbf{z}_2 - \frac{1}{2}\mathbf{z}_1 = \left(-\frac{1}{2}, -\frac{1}{2}, 3, 1\right)^T,$$

for $W^\perp$ is obtained by a single Gram-Schmidt step. To decompose the vector $\mathbf{v} = (1, 0, 0, 0)^T = \mathbf{w} + \mathbf{z}$, say, we compute the two orthogonal projections: Recall (3.57)

$$\mathbf{w} = \frac{1}{3}\mathbf{w}_1 + \frac{1}{7}\mathbf{w}_2 = \left(\frac{10}{21}, \frac{10}{21}, \frac{1}{7}, \frac{1}{21}\right)^T \in W$$
$$\mathbf{z} = \mathbf{v} - \mathbf{w} = -\frac{1}{2}\mathbf{y}_1 - \frac{1}{21}\mathbf{y}_2 = \left(\frac{11}{21}, -\frac{10}{21}, -\frac{1}{7}, -\frac{1}{21}\right)^T \in W^\perp. \qquad \blacksquare$$

3.4.5 Least Squares Minimization

For us, the most important case is that of a linear system

$$A\mathbf{x} = \mathbf{b} \tag{3.58}$$

consisting of m equations in n unknowns. In this case, the solutions may be obtained by minimizing the function

$$p(\mathbf{x}) = \|A\mathbf{x} - \mathbf{b}\|^2, \tag{3.59}$$

where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^m$. Clearly $p(\mathbf{x})$ has a minimum value of 0, which is achieved if and only if $\mathbf{x}$ is a solution to the linear system (3.58). Of course, it is not clear that we have gained much, since we already know how to solve $A\mathbf{x} = \mathbf{b}$ by Gaussian Elimination. However, this artifice turns out to have profound consequences.

Suppose that the linear system (3.58) does not have a solution, i.e., $\mathbf{b}$ does not lie in the image of the matrix A. This situation is very typical when there are more equations than unknowns. Such problems arise in data fitting, when the measured data points are all supposed to lie on a straight line, say, but rarely do so exactly, due to experimental error. Although we know there is no exact solution to the system, we might still try to find an approximate solution — a vector $\mathbf{x}^\star$ that comes as close to solving the system as possible.

One way to measure closeness is by looking at the magnitude of the error as measured by the residual vector $\mathbf{r} = \mathbf{b} - A\mathbf{x}$, i.e., the difference between the right- and left-hand sides of the system. The smaller its norm $\|\mathbf{r}\| = \|A\mathbf{x} - \mathbf{b}\|$, the better the attempted solution. For the Euclidean norm, the vector $\mathbf{x}^\star$ that minimizes the squared residual norm function (3.59) is known as the least squares solution to the linear system, because $\|\mathbf{r}\|^2 = r_1^2 + \cdots + r_n^2$ is the sum of the squares of the individual error components. As before, if the linear system (3.58) happens to have an actual solution, with $A\mathbf{x}^\star = \mathbf{b}$, then $\mathbf{x}^\star$ qualifies as the least squares solution too, since in this case, $\|A\mathbf{x}^\star - \mathbf{b}\| = 0$ achieves its absolute minimum. So least squares solutions include traditional solutions as special cases.

Unlike an exact solution, the least squares minimizer depends on the choice of inner product governing the norm; thus a suitable weighted norm can be introduced to emphasize or de-emphasize the various errors. While not the only possible approach, least squares is certainly the easiest to analyze and solve, and, hence, is often the method of choice for fitting functions to experimental data and performing statistical analysis.

The following minimization problem arises in elementary geometry, although its practical implications cut a much wider swath. Given a point $\mathbf{b} \in \mathbb{R}^m$ and a subset $V \subset \mathbb{R}^m$, find the

point $\mathbf{v}^\star \in V$ that is closest to $\mathbf{b}$. In other words, we seek to minimize the Euclidean distance $d(\mathbf{v}, \mathbf{b}) = \|\mathbf{v} - \mathbf{b}\|$ over all possible $\mathbf{v} \in V$.

The simplest situation occurs when V is a subspace of $\mathbb{R}^m$. In this case, the closest point problem can, in fact, be reformulated as a least squares minimization problem. Let $\mathbf{v}_1, \dots, \mathbf{v}_n$ be a basis for V. The general element $\mathbf{v} \in V$ is a linear combination of the basis vectors. We can write the subspace elements in the form

$$\mathbf{v} = x_1\mathbf{v}_1 + \cdots + x_n\mathbf{v}_n = A\mathbf{x},$$

where $A = (\mathbf{v}_1\mathbf{v}_2 \dots \mathbf{v}_n)$ is the $m \times n$ matrix formed by the (column) basis vectors and $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ are the coordinates of $\mathbf{v}$ relative to the chosen basis. In this manner, we can identify V with the image of A, i.e., the subspace spanned by its columns. Consequently, the closest point in V to $\mathbf{b}$ is found by minimizing $\|\mathbf{v} - \mathbf{b}\|^2 = \|A\mathbf{x} - \mathbf{b}\|^2$ over all possible $\mathbf{x} \in \mathbb{R}^n$. But this is exactly the same as the least squares function (3.59)! Thus, if $\mathbf{x}^\star$ is the least squares solution to the system $A\mathbf{x} = \mathbf{b}$, then $\mathbf{v}^\star = A\mathbf{x}^\star$ is the closest point to $\mathbf{b}$ belonging to $V = \operatorname{img} A$. In this way, we have established a profound and fertile connection between least squares solutions to linear systems and the geometrical problem of minimizing distances to subspaces. And, as we shall see, the closest point $\mathbf{v} \in V$ turns out to be the orthogonal projection of $\mathbf{b}$ onto the subspace.

The simplest algebraic equations are linear systems. As such, one must thoroughly understand them before venturing into the far more complicated nonlinear realm. For minimization problems, the starting point is the quadratic function. We shall see how the problem of minimizing a general quadratic function of n variables can be solved by linear algebra techniques.

Let us begin by reviewing the very simplest example: minimizing a scalar quadratic polynomial $p(x) = ax^2 + 2bx + c$ over all possible values of $x \in \mathbb{R}$. If $a > 0$, then the graph of p is a parabola opening upwards, and so there exists a unique minimum value. If $a < 0$, the parabola points downwards, and there is no minimum (although there is a maximum). If $a = 0$, the graph is a straight line, and there is neither minimum nor maximum over all $x \in \mathbb{R}$ - except in the trivial case $b = 0$ also, and the function $p(x) = c$ is constant, with every x qualifying as a minimum (and a maximum).

In the case $a > 0$, the minimum can be found by calculus. The critical points of a function, which are candidates for minima (and maxima), are found by setting its derivative to zero. In this case, differentiating, and solving $p'(x) = 2ax + 2b = 0$, we conclude that the only possible minimum value occurs at

$$x^\star = -\frac{b}{a}, \qquad \text{where} \qquad p\,(x^\star) = c - \frac{b^2}{a}. \tag{3.60}$$

Of course, one must check that this critical point is indeed a minimum, and not a maximum or inflection point. The second derivative test will show that $p''\,(x^\star) = 2a > 0$, and so $x^\star$ is at least a local minimum.

A more instructive approach to this problem — and one that requires only elementary algebra — is to "complete the square". We rewrite

$$p(x) = a\left(x + \frac{b}{a}\right)^2 + \frac{ac - b^2}{a}.$$

If $a > 0$, then the first term is always ≥ 0, and, moreover, attains its minimum value 0 only at $x^\star = -b/a$. The second term is constant, and so is unaffected by the value of x. Thus, the

global minimum of $p(x)$ is at $x^\star = -b/a$. Moreover, its minimal value equals the constant term, $p(x^\star) = c - b^2/a$, thereby reconfirming and strengthening the calculus result in (3.60).

Now that we have the one-variable case firmly in hand, let us turn our attention to the more substantial problem of minimizing quadratic functions of several variables. Thus, we seek to minimize a (real) quadratic polynomial

$$p(\mathbf{x}) = p(x_1, \dots, x_n) = \sum_{i,j=1}^{n} k_{ij} x_i x_j - 2 \sum_{i=1}^{n} f_i x_i + c, \tag{3.61}$$

depending on n variables $\mathbf{x} = (x_1, x_2, \dots, x_n)^T \in \mathbb{R}^n$. The coefficients k_{ij}, f_i and c are all assumed to be real. Moreover, we can assume, without loss of generality, that the coefficients of the quadratic terms are symmetric: $k_{ij} = k_{ji}$. Note that $p(\mathbf{x})$ is more general than a quadratic form in that it also contains linear and constant terms. We seek a global minimum, and so the variables $\mathbf{x}$ are allowed to vary over all of $\mathbb{R}^n$.

Let us begin by rewriting the quadratic function (3.61) in a more compact matrix notation:

$$p(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} - 2\mathbf{x}^T \mathbf{f} + c, \quad \mathbf{x} \in \mathbb{R}^n, \tag{3.62}$$

in which $K = (k_{ij})$ is a symmetric $n \times n$ matrix, $\mathbf{f} \in \mathbb{R}^n$ is a constant vector, and c is a constant scalar.

Theorem: If K is a positive definite (and hence symmetric) matrix, then the quadratic function (3.62) has a unique minimizer, which is the solution to the linear system

$$K\mathbf{x} = \mathbf{f}, \qquad \text{namely} \qquad \mathbf{x}^\star = K^{-1}\mathbf{f}. \tag{3.63}$$

The minimum value of $p(\mathbf{x})$ is equal to any of the following expressions:

$$p(\mathbf{x}^\star) = p\left(K^{-1}\mathbf{f}\right) = c - \mathbf{f}^T K^{-1}\mathbf{f} = c - \mathbf{f}^T\mathbf{x}^\star = c - (\mathbf{x}^\star)^T K \mathbf{x}^\star. \tag{3.64}$$

Proof: First recall that positive definiteness implies that K is a nonsingular matrix, and hence the linear system (3.63) has a unique solution $\mathbf{x}^\star = K^{-1}\mathbf{f}$. Then, for all $\mathbf{x} \in \mathbb{R}^n$, since $\mathbf{f} = K\mathbf{x}^\star$, it follows that

$$\begin{aligned} p(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} - 2\mathbf{x}^T \mathbf{f} + c &= \mathbf{x}^T K \mathbf{x} - 2\mathbf{x}^T K \mathbf{x}^\star + c \\ &= (\mathbf{x} - \mathbf{x}^\star)^T K (\mathbf{x} - \mathbf{x}^\star) + \left[c - (\mathbf{x}^\star)^T K \mathbf{x}^\star \right], \end{aligned} \tag{3.65}$$

where we used the symmetry of $K = K^T$ to identify the scalar terms

$$\mathbf{x}^T K \mathbf{x}^\star = \left(\mathbf{x}^T K \mathbf{x}^\star\right)^T = (\mathbf{x}^\star)^T K^T \mathbf{x} = (\mathbf{x}^\star)^T K \mathbf{x}.$$

The first term in the final expression in (3.65) has the form $\mathbf{y}^T K \mathbf{y}$, where $\mathbf{y} = \mathbf{x} - \mathbf{x}^\star$. Since we assumed that K is positive definite, we know that $\mathbf{y}^T K \mathbf{y} > 0$ for all $\mathbf{y} \neq \mathbf{0}$. Thus, the first term achieves its minimum value, namely 0, if and only if $\mathbf{0} = \mathbf{y} = \mathbf{x} - \mathbf{x}^\star$. Since $\mathbf{x}^\star$ is fixed, the second, bracketed, term does not depend on $\mathbf{x}$, and hence the minimizer of $p(\mathbf{x})$ coincides with the minimizer of the first term, namely $\mathbf{x} = \mathbf{x}^\star$. Moreover, the minimum value of $p(\mathbf{x})$ is equal to the constant term: $p(\mathbf{x}^\star) = c - (\mathbf{x}^\star)^T K \mathbf{x}^\star$. The alternative expressions in (3.64) follow from simple substitutions.

We are now ready to solve the geometric problem of finding the element in a prescribed subspace that lies closest to a given point. For simplicity, we work mostly with subspaces of $\mathbb{R}^m$, equipped with the Euclidean norm and inner product, but the method extends straightforwardly to arbitrary finite-dimensional subspaces of any inner product space. However, it does not apply to more general norms not associated with inner products, such as the 1 norm, the ∞ norm and, in fact, the p norms whenever $p \neq 2$. In such cases, finding the closest point problem is a nonlinear minimization problem whose solution requires more sophisticated analytical techniques.

Let $\mathbb{R}^m$ be equipped with an inner product $\langle \mathbf{v}, \mathbf{w} \rangle$ and associated norm $\|\mathbf{v}\|$, and let $W \subset \mathbb{R}^m$ be a subspace. Given $\mathbf{b} \in \mathbb{R}^m$, the goal is to find the point $\mathbf{w}^\star \in W$ that minimizes $\|\mathbf{w} - \mathbf{b}\|$ over all possible $\mathbf{w} \in W$. The minimal distance $d^\star = \|\mathbf{w}^\star - \mathbf{b}\|$ to the closest point is designated as the distance from the point $\mathbf{b}$ to the subspace W.

Of course, if $\mathbf{b} \in W$ lies in the subspace, then the answer is easy: the closest point in W is $\mathbf{w}^\star = \mathbf{b}$ itself, and the distance from $\mathbf{b}$ to the subspace is zero. Thus, the problem becomes interesting only when $\mathbf{b} \notin W$.

In solving the closest point problem, the goal is to minimize the squared distance

$$\|\mathbf{w} - \mathbf{b}\|^2 = \langle \mathbf{w} - \mathbf{b}, \mathbf{w} - \mathbf{b} \rangle = \|\mathbf{w}\|^2 - 2\langle \mathbf{w}, \mathbf{b} \rangle + \|\mathbf{b}\|^2 \tag{3.66}$$

over all possible $\mathbf{w}$ belonging to the subspace $W \subset \mathbb{R}^m$. Let us assume that we know a basis $\mathbf{w}_1, \dots, \mathbf{w}_n$ of W, with $n = \dim W$. Then the most general vector in W is a linear combination

$$\mathbf{w} = x_1 \mathbf{w}_1 + \cdots + x_n \mathbf{w}_n \tag{3.67}$$

of the basis vectors. We substitute the formula (3.67) for $\mathbf{w}$ into the squared distance function (3.66). As we shall see, the resulting expression is a quadratic function of the coefficients $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$, and so the minimum is provided by (3.63).

First, the quadratic terms come from expanding

$$\|\mathbf{w}\|^2 = \langle x_1 \mathbf{w}_1 + \cdots + x_n \mathbf{w}_n, x_1 \mathbf{w}_1 + \cdots + x_n \mathbf{w}_n \rangle = \sum_{i,j=1}^{n} x_i x_j \langle \mathbf{w}_i, \mathbf{w}_j \rangle .$$

Therefore,

$$\|\mathbf{w}\|^2 = \sum_{i,j=1}^{n} k_{ij} x_i x_j = \mathbf{x}^T K \mathbf{x}$$

where K is the symmetric $n \times n$ Gram matrix (3.39) whose (i, j) entry is the inner product

$$k_{ij} = \langle \mathbf{w}_i, \mathbf{w}_j \rangle \tag{3.68}$$

between the basis vectors of our subspace. Similarly,

$$\langle \mathbf{w}, \mathbf{b} \rangle = \langle x_1 \mathbf{w}_1 + \cdots + x_n \mathbf{w}_n, \mathbf{b} \rangle = \sum_{i=1}^{n} x_i \langle \mathbf{w}_i, \mathbf{b} \rangle ,$$

and so

$$\langle \mathbf{w}, \mathbf{b} \rangle = \sum_{i=1}^{n} x_i f_i = \mathbf{x}^T \mathbf{f},$$

where $\mathbf{f} \in \mathbb{R}^n$ is the vector whose ith entry is the inner product

$$f_i = \langle \mathbf{w}_i, \mathbf{b} \rangle \tag{3.69}$$

between the point and the subspace's basis elements. Substituting back, we conclude that the squared distance function (3.66) reduces to the quadratic function

$$p(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} - 2\mathbf{x}^T \mathbf{f} + c = \sum_{i,j=1}^{n} k_{ij} x_i x_j - 2 \sum_{i=1}^{n} f_i x_i + c,$$

in which K and $\mathbf{f}$ are given in (3.68) and (3.69), while $c = \|\mathbf{b}\|^2$. Since we assumed that the basis vectors $\mathbf{w}_1, \dots, \mathbf{w}_n$ are linearly independent, their associated Gram matrix is positive definite. Therefore, we may directly apply (3.63) to solve the closest point problem.

Theorem: Let $\mathbf{w}_1, \dots, \mathbf{w}_n$ form a basis for the subspace $W \subset \mathbb{R}^m$. Given $\mathbf{b} \in \mathbb{R}^m$, the closest point $\mathbf{w}^\star = x_1^\star \mathbf{w}_1 + \cdots + x_n^\star \mathbf{w}_n \in W$ is unique and prescribed by the solution $\mathbf{x}^\star = K^{-1}\mathbf{f}$ to the linear system

$$K\mathbf{x} = \mathbf{f}, \tag{3.70}$$

where the entries of K and $\mathbf{f}$ are given in (3.68) and (3.69). The (minimum) distance between the point and the subspace is

$$d^\star = \|\mathbf{w}^\star - \mathbf{b}\| = \sqrt{\|\mathbf{b}\|^2 - \mathbf{f}^T \mathbf{x}^\star}. \tag{3.71}$$

When the standard dot product and Euclidean norm on $\mathbb{R}^m$ are used to measure distance, the entries of the Gram matrix K and the vector $\mathbf{f}$ are given by

$$k_{ij} = \mathbf{w}_i \cdot \mathbf{w}_j = \mathbf{w}_i^T \mathbf{w}_j, \qquad f_i = \mathbf{w}_i \cdot \mathbf{b} = \mathbf{w}_i^T \mathbf{b}.$$

As in (3.41), each set of equations can be combined into a single matrix equation. If $A = \begin{pmatrix} \mathbf{w}_1 & \mathbf{w}_2 & \dots & \mathbf{w}_n \end{pmatrix}$ denotes the $m \times n$ matrix formed by the basis vectors, then

$$K = A^T A, \quad \mathbf{f} = A^T \mathbf{b}, \quad c = \mathbf{b}^T \mathbf{b} = \|\mathbf{b}\|^2. \tag{3.72}$$

A direct derivation of these equations is instructive. As

$$\mathbf{w} = x_1 \mathbf{w}_1 + \cdots + x_n \mathbf{w}_n = A\mathbf{x}$$

we have

$$\begin{aligned} \|\mathbf{w} - \mathbf{b}\|^2 = \|A\mathbf{x} - \mathbf{b}\|^2 &= (A\mathbf{x} - \mathbf{b})^T (A\mathbf{x} - \mathbf{b}) = \left(\mathbf{x}^T A^T - \mathbf{b}^T\right)(A\mathbf{x} - \mathbf{b}) \\ &= \mathbf{x}^T A^T A \mathbf{x} - 2\mathbf{x}^T A^T \mathbf{b} + \mathbf{b}^T \mathbf{b} = \mathbf{x}^T K \mathbf{x} - 2\mathbf{x}^T \mathbf{f} + c, \end{aligned}$$

thereby justifying (3.72). Thus, (3.70) and (3.71) imply that the closest point $\mathbf{w}^\star = A\mathbf{x}^\star \in W$ to $\mathbf{b}$ in the Euclidean norm is obtained by solving what are known as the *normal equations*

$$\left(A^T A\right)\mathbf{x} = A^T \mathbf{b} \tag{3.73}$$

for

$$\mathbf{x}^\star = \left(A^T A\right)^{-1} A^T \mathbf{b}, \qquad \text{giving} \qquad \mathbf{w}^\star = A\mathbf{x}^\star = A\left(A^T A\right)^{-1} A^T \mathbf{b}.$$

If, instead of the Euclidean inner product, we adopt a weighted inner product $\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T C \mathbf{w}$ on $\mathbb{R}^m$ prescribed by a positive definite $m \times m$ matrix $C > 0$, then the same computations produce

$$K = A^T C A, \quad \mathbf{f} = A^T C \mathbf{b}, \quad c = \mathbf{b}^T C \mathbf{b} = \|\mathbf{b}\|^2.$$

The resulting formula for the weighted Gram matrix K was previously derived in (3.43). In this case, the closest point $\mathbf{w}^\star \in W$ in the weighted norm is obtained by solving the *weighted normal equations*

$$A^T C A \mathbf{x} = A^T C \mathbf{b} \tag{3.74}$$

so that

$$\mathbf{x}^\star = \left(A^T C A\right)^{-1} A^T C \mathbf{b}, \quad \mathbf{w}^\star = A\mathbf{x}^\star = A\left(A^T C A\right)^{-1} A^T C \mathbf{b}. \tag{3.75}$$

Remark: The solution to the closest point problem given in (3.70) and (3.71) applies, as stated, to the more general case in which $W \subset V$ is a finite-dimensional subspace of a general inner product space V. The underlying inner product space V can even be infinite-dimensional, as, for example, in least squares approximations in function space.

Now, consider what happens if we know an orthonormal basis $\mathbf{u}_1, \dots, \mathbf{u}_n$ of the subspace W. Since, by definition, $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0$ for $i \neq j$, while $\langle \mathbf{u}_i, \mathbf{u}_i \rangle = \|\mathbf{u}_i\|^2 = 1$, the associated Gram matrix is the identity matrix: $K = I$. Thus, in this situation, the system (3.70) reduces to simply $\mathbf{x} = \mathbf{f}$, with solution $x_i^\star = f_i = \langle \mathbf{u}_i, \mathbf{b} \rangle$, and the closest point is given by

$$\mathbf{w}^\star = x_1^\star \mathbf{u}_1 + \cdots + x_n^\star \mathbf{u}_n \qquad \text{where} \qquad x_i^\star = \langle \mathbf{b}, \mathbf{u}_i \rangle, \quad i = 1, \dots, n. \tag{3.76}$$

We have already seen this formula! According to (3.56), $\mathbf{w}^\star$ is the orthogonal projection of $\mathbf{b}$ onto the subspace W. Thus, if we are supplied with an orthonormal basis of our subspace, we can easily compute the closest point using the orthogonal projection formula (3.76). If the basis is orthogonal, one can either normalize it or directly apply the equivalent orthogonal projection formula (3.57).

In this manner, we have established the key connection identifying the closest point in the subspace to a given vector with the orthogonal projection of that vector onto the subspace:

Theorem: Let $W \subset V$ be a finite-dimensional subspace of an inner product space. Given a point $\mathbf{b} \in V$, the closest point $\mathbf{w}^\star \in W$ coincides with the orthogonal projection of $\mathbf{b}$ onto W.

As we already observed, the solution to the closest point problem also solves the basic least squares minimization problem. Let us first officially define the notion of a (classical) least squares solution to a linear system.

Definition: A least squares solution to a linear system of equations

$$A\mathbf{x} = \mathbf{b} \tag{3.77}$$

is a vector $\mathbf{x}^\star \in \mathbb{R}^n$ that minimizes the squared Euclidean norm $\|A\mathbf{x} - \mathbf{b}\|^2$.

If the system (3.77) actually has a solution, then it is automatically the least squares solution. The concept of least squares solution is new only when the system does not have a solution, i.e., $\mathbf{b}$ does not lie in the image of A. We also want the least squares solution to be unique. As with an ordinary solution, this happens if and only if $\ker A = \{0\}$, or, equivalently, the columns of A are linearly independent, or, equivalently, $\operatorname{rank} A = n$. Indeed, if $\mathbf{z} \in \ker A$, then $\widetilde{\mathbf{x}} = \mathbf{x} + \mathbf{z}$ also satisfies

$$\|A\widetilde{\mathbf{x}} - \mathbf{b}\|^2 = \|A(\mathbf{x} + \mathbf{z}) - \mathbf{b}\|^2 = \|A\mathbf{x} - \mathbf{b}\|^2,$$

and hence is also a minimum. Thus, uniqueness requires $\mathbf{z} = \mathbf{0}$.

As before, to make the connection with the closest point problem, we identify the subspace $W = \operatorname{img} A \subset \mathbb{R}^m$ as the image or column space of the matrix A. If the columns of A are

linearly independent, then they form a basis for the image W. Since every element of the image can be written as $\mathbf{w} = A\mathbf{x}$, minimizing $\|A\mathbf{x} - \mathbf{b}\|^2$ is the same as minimizing the distance $\|\mathbf{w} - \mathbf{b}\|$ between the point and the subspace. The solution $\mathbf{x}^\star$ to the quadratic minimization problem produces the closest point $\mathbf{w}^\star = A\mathbf{x}^\star$ in $W = \operatorname{img} A$, which is thus found using (3.70) and (3.71). In the Euclidean case, we therefore find the least squares solution by solving the normal equations given in (3.73).

Theorem: Assume that $\ker A = \{0\}$. Then the least squares solution to the linear system $A\mathbf{x} = \mathbf{b}$ under the Euclidean norm is the unique solution $\mathbf{x}^\star$ to the normal equations

$$\left(A^T A\right)\mathbf{x} = A^T\mathbf{b}, \qquad \text{namely} \qquad \mathbf{x}^\star = \left(A^T A\right)^{-1} A^T\mathbf{b}. \tag{3.78}$$

The least squares error is

$$\|A\mathbf{x}^\star - \mathbf{b}\|^2 = \|\mathbf{b}\|^2 - \mathbf{f}^T\mathbf{x}^\star = \|\mathbf{b}\|^2 - \mathbf{b}^T A\left(A^T A\right)^{-1} A^T\mathbf{b}. \tag{3.79}$$

Note that the normal equations (3.73) can be simply obtained by multiplying the original system $A\mathbf{x} = \mathbf{b}$ on both sides by A^T. In particular, if A is square and invertible, then $\left(A^T A\right)^{-1} = A^{-1}\left(A^T\right)^{-1}$, and so the least squares solution formula (3.78) reduces to $\mathbf{x} = A^{-1}\mathbf{b}$, while the error formula (3.79) becomes zero. In the rectangular case — when inversion of A itself is not allowed — (3.78) gives a new formula for the solution to the linear system $A\mathbf{x} = \mathbf{b}$ whenever $\mathbf{b} \in \operatorname{img} A$. An alternative approach is to use a *pseudoinverse* of a matrix.

One can extend the basic least squares method by introducing a suitable weighted norm in the measurement of the error. Let $C > 0$ be a positive definite matrix that governs the weighted norm $\|\mathbf{v}\|^2 = \mathbf{v}^T C\mathbf{v}$. In most applications, $C = \operatorname{diag}(c_1, \dots, c_m)$ is a diagonal matrix whose entries are the assigned weights of the individual coordinates, but the method works equally well for general norms defined by positive definite matrices. The off-diagonal entries of C can be used to weight cross-correlations between data values, although this extra freedom is rarely used in practice.[26] The weighted least squares solution is thus obtained by solving the corresponding weighted normal equations (3.74), as follows.

Theorem: Suppose A is an $m \times n$ matrix such that $\ker A = \{\mathbf{0}\}$, and suppose $C > 0$ is any positive definite $m \times m$ matrix specifying the weighted norm $\|\mathbf{v}\|^2 = \mathbf{v}^T C\mathbf{v}$. Then the least squares solution to the linear system $A\mathbf{x} = \mathbf{b}$ that minimizes the weighted squared error $\|A\mathbf{x} - \mathbf{b}\|^2$ is the unique solution $\mathbf{x}^\star$ to the weighted normal equations

$$A^T C A\mathbf{x}^\star = A^T C\mathbf{b}, \qquad \text{so that} \qquad \mathbf{x}^\star = \left(A^T C A\right)^{-1} A^T C\mathbf{b}.$$

The weighted least squares error is

$$\|A\mathbf{x}^\star - \mathbf{b}\|^2 = \|\mathbf{b}\|^2 - \mathbf{f}^T\mathbf{x}^\star = \|\mathbf{b}\|^2 - \mathbf{b}^T C A\left(A^T A\right)^{-1} A^T C\mathbf{b}.$$

[26] I could not disagree with the authors more on this point. Regression with autocorrelated errors is one of the most fundamental core topics in time-series econometrics. See, e.g., Paolella, Linear Models and Time-Series Analysis, chapters 4 to 9.

4 Multivariate Calculus: Differentiation, Tangent Maps, and Taylor Series

> The discovery in 1846 of the planet Neptune was a dramatic and spectacular achievement of mathematical astronomy. The very existence of this new member of the solar system, and its exact location, were demonstrated with pencil and paper; there was left to observers only the routine task of pointing their telescopes at the spot the mathematicians had marked. (James R. Newman)
>
> I was brought up to believe The universe has a plan We are only human. It's not ours to understand (Rush, BU2B)

4.1 Sequences, Limits, Functions, and Continuity

> Compactness is the single most important concept in real analysis. It is what reduces the infinite to the finite.
>
> (Charles Pugh, Real Mathematical Analysis, 2nd edition, 2015, p. 79)

This short section states some key definitions we will require throughout the remainder of the document. The emphasis is on multivariate sequences and continuity of multivariate functions. The above quote from Pugh regarding compactness has merit, notably in functional analysis, though is arguably a bit exaggerated in our context: There is a solid handful of core concepts, such as the ones stated below, that are yet more crucial for (univariate and multivariate) analysis. In this document, we will not require the use of compactness, though it does lurk in the background, namely in the proofs of some results that we omit.

- For any $\mathbf{x} \in \mathbb{R}^n$ and $r \in \mathbb{R}_{>0}$, the *open ball of radius r around* $\mathbf{x}$ is the subset $B_r(\mathbf{x}) \subset \mathbb{R}^n$ with $B_r(\mathbf{x}) := \{\mathbf{y} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{y}\| < r\}$ (note the strict inequality), where, recalling (1.11), $\|\mathbf{x}\|$ is the *norm* of $\mathbf{x}$. We will also use the calligraphic B, i.e., $\mathcal{B}_r(\mathbf{x})$.
- A *neighborhood* of a point $\mathbf{x} \in \mathbb{R}^n$ is a subset $A \subset \mathbb{R}^n$ such that there exists an $\epsilon > 0$ with $B_\epsilon(\mathbf{x}) \subset A$.
- If, for some $r \in \mathbb{R}_{>0}$, the set $A \subset \mathbb{R}^n$ is contained in the ball $B_r(\mathbf{0})$, then A is said to be *bounded*.
- The subset $U \subset \mathbb{R}^n$ is *open in* $\mathbb{R}^n$ if, for every point $\mathbf{x} \in U$, $\exists r > 0$ such that $B_r(\mathbf{x}) \subset U$.

  We prove below in (4.1) that, for every $\mathbf{x} \in \mathbb{R}^n$ and $r > 0$, the open ball $B_r(\mathbf{x})$ is open in $\mathbb{R}^n$. For example, the open interval $\{x \in \mathbb{R} : |x - c| < r\} = (c - r, c + r)$, $c \in \mathbb{R}$, $r \in \mathbb{R}_{>0}$, is an open set in $\mathbb{R}$, but it is not open in the plane $\mathbb{R}^2$. Likewise, the square region $S_1 = \{\mathbf{y} = (y_1, y_2) \in \mathbb{R}^2 : |y_1| < 1, |y_2| < 1\}$ is open in $\mathbb{R}^2$, but not in $\mathbb{R}^3$.
- A set $C \subset \mathbb{R}^n$ is *closed* if its complement, $\mathbb{R}^n \setminus C$ is open. By convention, the empty set $\emptyset$ is open (indeed, every point in $\emptyset$ satisfies the requirement), so that its complement, $\mathbb{R}^n$, is closed. But, from the definition, $\mathbb{R}^n$ is open, so that $\emptyset$ is closed. This is not incorrect: sets can be open and closed (or neither). The closed interval $[a, b]$ is a closed set, as is the square region $S_2 = \{\mathbf{y} = (y_1, y_2) \in \mathbb{R}^2 : |y_1| \leq 1, |y_2| \leq 1\}$.

  Below, after vector sequences are introduced, we state a definition of a closed set that is equivalent to its above definition in terms of open sets, but adds considerably more intuition into what a closed set represents.

- The point $\mathbf{x} \in A \subset \mathbb{R}^n$ is an *interior point* of A if $\exists r > 0$ such that $B_r(\mathbf{x}) \subset A$.
- The *interior* of A is the set of all interior points of A, denoted A^o. Observe that the biggest open set contained in any set $A \subset \mathbb{R}^n$ is A^o.
- The smallest closed set that contains A is the *closure* of A, denoted $\overline{A}$; it is the set of $\mathbf{x} \in \mathbb{R}^n$ such that, $\forall r > 0$, $B_r(\mathbf{x}) \cap A \neq \emptyset$.

  The closure of a set is closed. For example, the closure of (a, b), $a < b$, is $[a, b]$ (note that its complement, $(-\infty, a) \cup (b, \infty)$ is open).

  As an example in $\mathbb{R}^2$, using the sets S_1 and S_2 given above, $\overline{S}_1 = S_2$. In words, the closure of a set includes its "boundary"; see the next item.
- The *boundary* of a set $A \subset \mathbb{R}^n$, denoted ∂A, is defined to be the difference between its closure and interior, i.e., $\partial A = \overline{A} - A^o$. (The notation ∂ is used because it signifies a line around a region, and has nothing to do with the symbol for the partial derivative.)

Example 4.1 *For points $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$, the* line segment *from $\mathbf{x}_1$ to $\mathbf{x}_2$ is the set of points*

$$\mathbf{x}_1 + t(\mathbf{x}_2 - \mathbf{x}_1) = t\mathbf{x}_2 + (1-t)\mathbf{x}_1, \quad 0 \leq t \leq 1.$$

For point $\mathbf{c} \in \mathbb{R}^n$ and $r > 0$, let $B_r(\mathbf{c})$ be the open ball of radius r around $\mathbf{c}$. It should be geometrically obvious that, if $\mathbf{x}_1, \mathbf{x}_2 \in B_r(\mathbf{c})$, then so are all the points on the line segment from $\mathbf{x}_1$ to $\mathbf{x}_2$. To see this algebraically, let $\mathbf{x} = \mathbf{x}(t) = t\mathbf{x}_2 + (1-t)\mathbf{x}_1$ for $0 \leq t \leq 1$, and use the triangle inequality (1.10) to get

$$\begin{aligned} \|\mathbf{x} - \mathbf{c}\| &= \|t\mathbf{x}_2 + (1-t)\mathbf{x}_1 - t\mathbf{c} - (1-t)\mathbf{c}\| \\ &= \|t(\mathbf{x}_2 - \mathbf{c}) + (1-t)(\mathbf{x}_1 - \mathbf{c})\| \\ &\leq t\|\mathbf{x}_2 - \mathbf{c}\| + (1-t)\|\mathbf{x}_1 - \mathbf{c}\| \\ &< t \cdot r + (1-t) \cdot r = r. \end{aligned}$$

As $B_r(\mathbf{c})$ is open, $\|\mathbf{x} - \mathbf{c}\|$ is strictly less than r, though $\sup \|\mathbf{x} - \mathbf{c}\| = r$. ■

The following presentation of the above mentioned result is from Fitzpatrick, p. 284.

Theorem:

Every open ball in $\mathbb{R}^n$ is open in $\mathbb{R}^n$. (4.1)

Proof: Let $\mathbf{u}$ be a point in $\mathbb{R}^n$ and let r be a positive real number. Consider the open ball $\mathcal{B}_r(\mathbf{u})$. We must show that every point in $\mathcal{B}_r(\mathbf{u})$ is an interior point of $\mathcal{B}_r(\mathbf{u})$. See Figure 30. Let $\mathbf{v}$ be a point in $\mathcal{B}_r(\mathbf{u})$. Define $R = r - \text{dist}(\mathbf{u}, \mathbf{v})$ and observe that R is positive. We claim that

$$\mathcal{B}_R(\mathbf{v}) \subseteq \mathcal{B}_r(\mathbf{u}). \tag{4.2}$$

Indeed, if $\mathbf{w}$ is in $\mathcal{B}_R(\mathbf{v})$, then $\text{dist}(\mathbf{w}, \mathbf{v}) < R = r - \text{dist}(\mathbf{u}, \mathbf{v})$, so, using the triangle inequality, we have

$$\begin{aligned} \text{dist}(\mathbf{w}, \mathbf{u}) &\leq \text{dist}(\mathbf{w}, \mathbf{v}) + \text{dist}(\mathbf{v}, \mathbf{u}) \\ &< [r - \text{dist}(\mathbf{u}, \mathbf{v})] + \text{dist}(\mathbf{v}, \mathbf{u}) = r. \end{aligned}$$

Thus, the inclusion (4.2) holds; so $\mathbf{v}$ is an interior point of $\mathcal{B}_r(\mathbf{u})$.

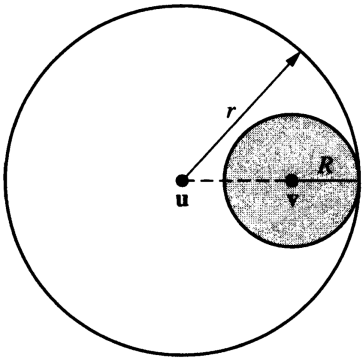


Figure 30: From Fitzpatrick, p. 284. $R = r - \text{dist}(\mathbf{u}, \mathbf{v})$. $\mathcal{B}_R(\mathbf{v}) \subseteq \mathcal{B}_r(\mathbf{u})$ if $R = r - \text{dist}(\mathbf{u}, \mathbf{v})$.

The next result (4.3) offers further practice with the above definitions, and we will need it directly after to prove (4.4), which in turn is used to prove result (4.138). The following two theorems are standard results; I took them from Sasane, A Friendly Approach to Functional Analysis, 2017, pp. 268, 269. The reference to "normed space", for our purposes, can be taken to be the usual vector space in $\mathbb{R}^n$, with the usual vector norm; see the definition in §3.4.1.

Theorem: Let $(X, \|\cdot\|)$ be a normed space, $x \in X$ and $r > 0$. Prove that

$$\overline{B(x,r)} := \{y \in X : \|y - x\| \leqslant r\} \textbf{ is a closed set.} \tag{4.3}$$

Proof: Consider the closed ball $\overline{B(x,r)} = \{y \in X : \|x - y\| \leqslant r\}$ in X. To show that $\overline{B(x,r)}$ is closed, we'll show its complement, $U := \{y \in X : \|x - y\| > r\}$, open. If $y \in U$, then $\|x - y\| > r$. Define $r' = \|x - y\| - r > 0$. We claim that $B(y, r') \subset U$. Let $z \in B(y, r')$. Then $\|z - y\| < r' = \|x - y\| - r$ and so, from the reverse triangle (1.8) with $a = x - y$ and $b = y - z$,

$$\|x - z\| \geqslant \|x - y\| - \|y - z\| > \|x - y\| - (\|x - y\| - r) = r.$$

Hence $z \in U$.

Theorem: The unit sphere is closed, i.e.,

$$\mathbb{S} := \{\mathbf{x} \in X : \|x\| = 1\} \textbf{ is closed.} \tag{4.4}$$

Proof: We know from (4.1) that the interior of $\mathbb{S}$, namely the open ball $B(\mathbf{0}, 1) = \{\mathbf{x} \in X : \|\mathbf{x}\| < 1\}$ is open. Also, it follows from (4.3) that the exterior of the closed ball $\overline{B(\mathbf{0}, 1)}$, namely the set $U = \{\mathbf{x} \in X : \|\mathbf{x}\| > 1\}$ is open as well. Thus, the complement of $\mathbb{S}$, being the union of the two open sets $B(\mathbf{0}, 1)$ and U, is open. Consequently, $\mathbb{S}$ is closed.

We continue now with essential concepts and results.

- A (multivariate, or vector) sequence is a *mapping* $\mathbf{f} : \mathbb{N} \to \mathbb{R}^n$ with kth term $\mathbf{f}(k)$, $k \in \mathbb{N}$. Many authors reserve the word *function* for the case when $n = 1$. As in the univariate case, the more common notation for sequence $\mathbf{a}_1 = \mathbf{f}(1)$, $\mathbf{a}_2 = \mathbf{f}(2), \ldots$ is $\{\mathbf{a}_k\}$. The ith component of $\mathbf{a}_k$ is denoted by $(\mathbf{a}_k)_i$, $i = 1, \ldots, n$. For sequence $\{\mathbf{a}_k\}$ and set $S \subset \mathbb{R}^n$, the notation $\{\mathbf{a}_k\} \in S$ indicates that, $\forall k \in \mathbb{N}$, $\mathbf{a}_k \in S$.

- The sequence $\{\mathbf{a}_k\} \in \mathbb{R}^n$ converges to $\mathbf{a} \in \mathbb{R}^n$ if, $\forall \epsilon > 0$, $\exists K \in \mathbb{N}$ such that, $\forall k > K$, $\|\mathbf{a}_k - \mathbf{a}\| < \epsilon$. Point $\mathbf{a}$ is the *limit* of $\{\mathbf{a}_k\}$ if $\{\mathbf{a}_k\}$ converges to $\mathbf{a}$, in which case one writes $\lim_{k\to\infty} \mathbf{a}_k = \mathbf{a}$. As in the univariate case, if the limit exists, then it is unique.

- In order for $\lim_{k\to\infty} \mathbf{a}_k = \mathbf{a} = (a_1, \ldots, a_n)$ to hold, it is necessary and sufficient that $\lim_{k\to\infty} (\mathbf{a}_k)_i = a_i$, $i = 1, \ldots, n$.

  Definition: For each index i with $1 \le i \le n$, we define the ith component projection function $p_i : \mathbb{R}^n \to \mathbb{R}$ by

$$p_i(\mathbf{u}) \equiv u_i, \qquad \text{for } \mathbf{u} = (u_1, \ldots, u_n) \in \mathbb{R}^n. \tag{4.5}$$

  It follows directly from this definition that, for $\mathbf{u} \in \mathbb{R}^n$, $\mathbf{u} = (p_1(\mathbf{u}), \ldots, p_n(\mathbf{u}))$, so a point in $\mathbb{R}^n$ is completely determined by the values of the component projection functions at that point.

  Definition: A sequence of points $\{\mathbf{u}_k\}$ in $\mathbb{R}^n$ is said to converge componentwise to the point $\mathbf{u}$ in $\mathbb{R}^n$ provided that, for each index i with $1 \le i \le n$,

$$\lim_{k\to\infty} p_i(\mathbf{u}_k) = p_i(\mathbf{u}).$$

  Theorem (The Componentwise Convergence Criterion): Let $\{\mathbf{u}_k\}$ be a sequence in $\mathbb{R}^n$ and let $\mathbf{u}$ be a point in $\mathbb{R}^n$. Then

$$\{\mathbf{u}_k\} \to \mathbf{u} \iff \{\mathbf{u}_k\} \textbf{ converges componentwise to } \mathbf{u}. \tag{4.6}$$

  Proof: First we suppose that the sequence $\{\mathbf{u}_k\}$ converges to $\mathbf{u}$. Fix an index i with $1 \le i \le n$. Then

$$0 \le |p_i(\mathbf{u}_k) - p_i(\mathbf{u})| = |p_i(\mathbf{u}_k - \mathbf{u})| \le \|\mathbf{u}_k - \mathbf{u}\| \qquad \text{for every index } k.$$

  Since, by definition, the sequence of real numbers $\{\|\mathbf{u}_k - \mathbf{u}\|\}$ converges to 0, it follows from (2.15) that

$$0 \le \lim_{k\to\infty} |p_i(\mathbf{u}_k) - p_i(\mathbf{u})| \le \lim_{k\to\infty} \|\mathbf{u}_k - \mathbf{u}\| = 0;$$

  that is, the sequence $\{p_i(\mathbf{u}_k)\}$ converges to $p_i(\mathbf{u})$. Thus, $\{\mathbf{u}_k\}$ converges componentwise to $\mathbf{u}$.

  To prove the converse, suppose that the sequence $\{\mathbf{u}_k\}$ converges componentwise to $\mathbf{u}$. Then, by definition, for each index i with $1 \le i \le n$, $\lim_{k\to\infty} p_i(\mathbf{u}_k - \mathbf{u}) = 0$. But then by the addition and product properties of convergent real sequences, namely (2.12) and (2.13), respectively, it follows that

$$\lim_{k\to\infty} \left[(p_1(\mathbf{u}_k - \mathbf{u}))^2 + \cdots + (p_n(\mathbf{u}_k - \mathbf{u}))^2\right] = 0.$$

  This last assertion means precisely that $\lim_{k\to\infty} \|\mathbf{u}_k - \mathbf{u}\|^2 = 0$, and hence, by the continuity of the square root function and (2.19), $\lim_{k\to\infty} \|\mathbf{u}_k - \mathbf{u}\| = 0$; that is, the sequence $\{\mathbf{u}_k\}$ converges to $\mathbf{u}$.

- The point $\mathbf{a} \in S \subset \mathbb{R}^n$ is said to be a *limit point* of S if $\exists \{\mathbf{a}_k\} \in S$ such that $\lim_{k\to\infty} \mathbf{a}_k = \mathbf{a}$. In other words, $\mathbf{a}$ is a limit point of S if there exists a sequence with terms in S that converge to it.[27]

- As in the univariate case, $\{\mathbf{a}_k\}$ is a *Cauchy sequence* if, for a given $\epsilon > 0$, $\exists N \in \mathbb{N}$ such that $\forall n, m \geq N$, $\|\mathbf{a}_m - \mathbf{a}_n\| < \epsilon$. As expected, (2.198) generalizes: sequence $\{\mathbf{a}_k\}$ converges iff $\{\mathbf{a}_k\}$ is a Cauchy sequence.

- As in the univariate case, the series $\sum_{k=1}^{\infty} \mathbf{a}_k$ is convergent if the sequence of partial sums, $\{\mathbf{s}_p\}$, where $\mathbf{s}_p = \sum_{k=1}^{p} \mathbf{a}_k$, is convergent.

- Consider the function $f : \mathbb{N} \to I$, $I = (0, 1]$, given by $f(k) = 1/k$. Observe that I is neither open nor closed. Clearly, $\lim_{k\to\infty} a_k = 0$, and $0 \notin I$. However, $\lim_{k\to\infty} a_k$ is contained in the closure of I, which is the closed set $[0, 1]$. With this concept in mind, the following basic result of analysis should appear quite reasonable: The set $C \subset \mathbb{R}^n$ is *closed* iff it contains all its limit points.

- Given a mapping $F : A \to \mathbb{R}^m$, $A \subset \mathbb{R}^n$, and an index i, $1 \leq i \leq m$, we define the function $F_i : A \to \mathbb{R}$ to be the composition of $F : A \to \mathbb{R}^m$ with the ith *component projection*, where the latter is the function from A to $\mathbb{R}$ that returns the ith value of F, $1 \leq i \leq m$. We call the function $F_i : A \to \mathbb{R}$ the ith component function of the mapping $F : A \to \mathbb{R}^m$. Thus,

$$F(\mathbf{u}) = (F_1(\mathbf{u}), \ldots, F_m(\mathbf{u})), \quad \text{for } \mathbf{u} \in A, \tag{4.7}$$

and the mapping $F : A \to \mathbb{R}^m$ is said to be represented by its component functions as

$$F = (F_1, \ldots, F_m) : A \to \mathbb{R}^m. \tag{4.8}$$

For example, let $\mathcal{O}$ be the set of all nonzero points in $\mathbb{R}^n$. Define the mapping $F : \mathcal{O} \to \mathbb{R}^n$ by $F(\mathbf{u}) = \mathbf{u}/\|\mathbf{u}\|^2$, $\mathbf{u} \in \mathcal{O}$. Then the representation of the mapping in component functions is

$$F(\mathbf{u}) = \left(u_1/\|\mathbf{u}\|^2, \ldots, u_n/\|\mathbf{u}\|^2\right), \quad \mathbf{u} \in \mathcal{O}.$$

For $n = 3$, this component representation can be written as

$$F(x, y, z) = \left(\frac{x}{x^2 + y^2 + z^2}, \frac{y}{x^2 + y^2 + z^2}, \frac{z}{x^2 + y^2 + z^2}\right), \quad (x, y, z) \in \mathcal{O}.$$

- We now turn to limits of multivariate functions. Let $\mathbf{f} : A \subset \mathbb{R}^n \to \mathbb{R}^m$ be a mapping, and let $\mathbf{x}_0 \in \overline{A}$ be a point in the closure of A. Then $\lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{f}(\mathbf{x}) = \mathbf{b}$ if, $\forall \epsilon > 0$, $\exists \delta > 0$ such that

$$\|\mathbf{x} - \mathbf{x}_0\| < \delta, \;\; \mathbf{x} \in A \implies \|\mathbf{f}(\mathbf{x}) - \mathbf{b}\| < \epsilon. \tag{4.9}$$

We can state this limit result also in terms of sequences. We do so for $m = 1$, and allow the reader to state the general m case. We take the result from Fitzpatrick, Advanced Calculus (2009, Thm 13.7), who uses the notation $\text{dist}(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|$, and state the result for the $m = 1$ case.

[27] For this definition, some mathematicians will additionally require that the sequence consists of unique terms, i.e., $\mathbf{a}_k \in S$ with $\mathbf{a}_k \neq \mathbf{a}_h$, $h \neq k$, which, for example, precludes a finite set of points from having limit points. In this case, as in Stoll (2001, p. 69), a point $\mathbf{a} \in \mathbb{R}^n$ is a limit point of S if, $\forall r > 0$, $(B_r(\mathbf{a}) \setminus \mathbf{a}) \cap S \neq \emptyset$. In words, $\mathbf{a}$ is a limit point of S if, for any $r > 0$, the open ball of radius r around $\mathbf{a}$, but "punctured" at $\mathbf{a}$ (i.e., excluding the point $\mathbf{a}$) contains at least one point that is in S. Pugh (2002, p. 58) and Hubbard and Hubbard (2002, p. 97) are examples of authors that do not require $\mathbf{a}_k \neq \mathbf{a}_h$.

For $A \subset \mathbb{R}^n$, let $\mathbf{x}_*$ be a limit point of A. For a function $f : A \to \mathbb{R}$ and $\ell \in \mathbb{R}$, the following two assertions are equivalent:

i. $\lim_{\mathbf{x}\to\mathbf{x}_*} f(\mathbf{x}) = \ell$. That is, $\forall \{\mathbf{x}_k\} \in A \setminus \{\mathbf{x}_*\}$,

$$\lim_{k\to\infty} \mathbf{x}_k = \mathbf{x}_* \quad \Longrightarrow \quad \lim_{k\to\infty} f(\mathbf{x}_k) = \ell. \tag{4.10}$$

ii. $\forall \epsilon > 0$, $\exists \delta$ such that

$$|f(\mathbf{x}) - \ell| < \epsilon \quad \text{if } \mathbf{x} \text{ is in } A \setminus \{\mathbf{x}_*\} \text{ and } \operatorname{dist}(\mathbf{x}, \mathbf{x}_*) < \delta. \tag{4.11}$$

- Analogous to (2.12), for $f, g : A \subset \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{x}_0 \in \overline{A}$, if $\lim_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x})$ and $\lim_{\mathbf{x}\to\mathbf{x}_0} g(\mathbf{x})$ exist, then, for $k_1, k_2 \in \mathbb{R}$,

$$\lim_{\mathbf{x}\to\mathbf{x}_0} \big(k_1 f(\mathbf{x}) + k_2 g(\mathbf{x})\big) = k_1 \lim_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x}) + k_2 \lim_{\mathbf{x}\to\mathbf{x}_0} g(\mathbf{x}). \tag{4.12}$$

- Just as we have the Componentwise Convergence Criterion for the convergence of sequences in Euclidean space, we also have the following simple, useful criterion for the continuity of a mapping.

  Theorem (The Componentwise Continuity Criterion): Let A be a subset of $\mathbb{R}^n$ that contains the point $\mathbf{u}$ and consider the mapping

$$F = (F_1, \ldots, F_m) : A \to \mathbb{R}^m.$$

  Then the mapping $F : A \to \mathbb{R}^m$ is continuous at $\mathbf{u}$ if and only if each of its component functions $F_i : A \to \mathbb{R}$ is continuous at $\mathbf{u}$. In short,

$$F : A \to \mathbb{R}^m \textbf{ is continuous at } \mathbf{u} \iff F_i : A \to \mathbb{R} \textbf{ is continuous at } \mathbf{u}. \tag{4.13}$$

Proof: This result follows immediately from the Componentwise Convergence Criterion (4.6) since, if $\{\mathbf{u}_k\}$ is a sequence in A that converges to the point $\mathbf{u}$, then the image sequence $\{F(\mathbf{u}_k)\}$ converges to $F(\mathbf{u})$ if and only if for each index i with $1 \le i \le m$, the sequence $\{F_i(\mathbf{u}_k)\}$ converges to $F_i(\mathbf{u})$.

Example 4.2 *Let $f(x) = 1/x$ for $x \in D = \mathbb{R} \setminus \{0\}$. It is easy to see that $\lim_{x\to 0} f(x)$ does not exist, though one-sided limits do exist in $\mathbb{X}$.* ■

Similar phenomena exist in the multivariate case. The next example, perhaps common to every textbook, illustrates an idea of how to show that the limit at a particular point does *not* exist. More work and different ideas, to prove that the limit exists, are required. This will be explored in Examples 4.5 and 4.6.

Example 4.3 *Let $f : A \to \mathbb{R}$ with $A = \mathbb{R}^2 \setminus \mathbf{0}$ and $f(\mathbf{x}) = x_1x_2/(x_1^2 + x_2^2)$. To see that $\lim_{\mathbf{x}\to\mathbf{0}} f(\mathbf{x})$ does not exist, set $x_2(x_1) = kx_1$ for some fixed $k \in \mathbb{R}$ so that $\lim_{x_1\to 0} x_2(x_1) = 0$ and $f(\mathbf{x}) = f(x_1, x_2(x_1)) = f(x_1, kx_1) = kx_1^2/(x_1^2 + k^2x_1^2) = k/(1 + k^2)$. Thus, along the line $x_2 = kx_1$, $\lim_{\mathbf{x}\to\mathbf{0}} f(\mathbf{x}) = k/(1 + k^2)$, i.e., it depends on the choice of k, showing that $\lim_{\mathbf{x}\to\mathbf{0}} f(\mathbf{x})$ depends on the path that $\mathbf{x}$ takes towards zero. Thus, $\lim_{\mathbf{x}\to\mathbf{0}} f(\mathbf{x})$ cannot exist.*

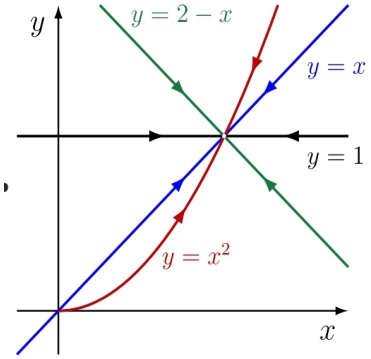
Another way to see this is to first observe that the sequence $\{(1/k, 1/k)\}$ *converges to the point* $(0,0)$*; and as* $f(1/k, 1/k) = 1/2$ *for each* $k \in \mathbb{N}$*, it follows that the image sequence* $\{f(1/k, 1/k)\}$ *converges to* $1/2$*. On the other hand, the sequence* $\{(1/k, 0)\}$ *also converges to the point* $(0,0)$*, and as* $f(1/k, 0) = 0$ *for each* $k \in \mathbb{N}$*, it follows that the image sequence* $\{f(1/k, 0)\}$ *converges to 0. Thus,* $\lim_{\mathbf{x}\to\mathbf{0}} f(\mathbf{x})$ *does not exist.* ■

Example 4.4 *As in Miklavcic, An Illustrative Guide to Multivariable and Vector Calculus (2020, p. 56), consider*

$$\lim_{(x,y)\to(1,1)} \frac{x-y}{x-1}.$$

This is of the form $0/0$*, so the function is undefined at* $(1,1)$*. Evaluating the limit by approaching the point* $(1,1)$ *along four different paths, we obtain*

$$\begin{aligned}
y = x &: \lim_{(1,1)} \frac{x-y}{x-1} = \lim_{(1,1)} \frac{0}{x-1} = 0. \\
y = 2-x &: \lim_{(1,1)} \frac{x-y}{x-1} = \lim_{(1,1)} \frac{2(x-1)}{x-1} = 2. \\
y = x^2 &: \lim_{(1,1)} \frac{x-y}{x-1} = \lim_{(1,1)} \frac{x-x^2}{x-1} = -1. \\
y = 1 &: \lim_{(1,1)} \frac{x-y}{x-1} = \lim_{(1,1)} \frac{x-1}{x-1} = 1,
\end{aligned}$$



as shown in the accompanying figure. The fact that at least two paths result in a different limit indicate that the function is not continuous at $(1,1)$*.* ■

Example 4.5 *Consider the following limit:*

$$\lim_{(x,y)\to(0,0)} \frac{x^3}{x^2+y^2}.$$

Define $f(x,y) = x^3/(x^2+y^2)$ *if* $(x,y) \neq (0,0)$*. By observing that, for* $k \in \mathbb{N}$*, the sequence* $\{(0, 1/k)\}$ *converges to* $(0,0)$ *and that* $f(0, 1/k) = 0$ *for each index* k*, we see that the only possible value of the limit is 0. To verify that the limit is indeed 0, it is necessary to make some estimates of the size of* $f(x,y)$*. Indeed, if* $x \neq 0$*, then*

$$\left|\frac{x^3}{x^2+y^2}\right| \leq \left|\frac{x^3}{x^2}\right| = |x|,$$

and, therefore,

$$\left|\frac{x^3}{x^2+y^2}\right| \leq |x| \quad \text{if } (x,y) \neq (0,0),$$

since this estimate also clearly holds if $x = 0$ *and* $y \neq 0$*. Now suppose that the sequence* $\{(x_k, y_k)\}$ *converges to* $(0,0)$ *with each* $(x_k, y_k) \neq (0,0)$*. Then the sequence* $\{x_k\}$ *converges to 0, so from the preceding estimate and the comparison test for convergent sequences* (2.224)*, it follows that the image sequence* $\{f(x_k, y_k)\}$ *converges to 0. Thus,*

$$\lim_{(x,y)\to(0,0)} \frac{x^3}{x^2+y^2} = 0.$$

More generally, let m *and* n *be natural numbers. One can show that the limit*

$$\lim_{(x,y)\to(0,0)} \frac{x^n y^m}{x^2+y^2}$$

exists if and only if $m+n > 2$*.* ■

Example 4.6 *As in Miklavcic (2020, p. 58), consider*

$$\lim_{(x,y)\to(0,0)} \frac{x^3 - x^2y}{x^2 + y^2 + xy}.$$

Note that the function is undefined at the origin. First we evaluate the limit along a few simple paths. Along $y = 0$ and $x = 0$, respectively, we have

$$\lim_{(0,0)} \frac{x^3 - x^2y}{x^2 + y^2 + xy} = \lim_{(0,0)} \frac{x^3}{x^2} = \lim_{x\to 0} x = 0; \quad \lim_{(0,0)} \frac{x^3 - x^2y}{x^2 + y^2 + xy} = \lim_{y\to 0} \frac{0}{y^2} = 0.$$

We get the same result along any straight line $y = kx$. If the limit exists, it must be 0. This analysis with straight lines is not adequate to declare that the limit exists. Consider an arbitrary curve $r = f(\theta) > 0$, where $x = r\cos\theta$, $y = r\sin\theta$, and let $r \to 0$. Figure 31 shows a typical case.

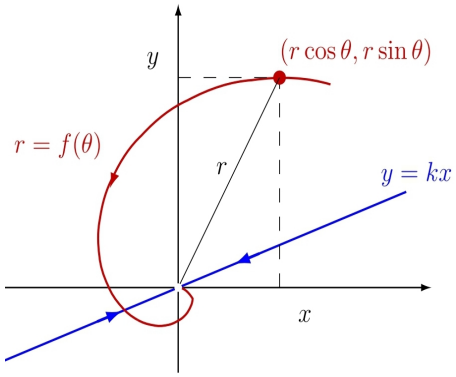


Figure 31: Spiral paths all lead to 0

Substitute the polar functions for x and y in the definition of the function limit to get

$$\left|\frac{x^3 - x^2y}{x^2 + y^2 + xy} - 0\right| = \left|\frac{r^3\cos^3\theta - r^3\cos^2\theta\sin\theta}{r^2\cos^2\theta + r^2\cos^2\theta + r^2\cos\theta\sin\theta}\right|$$
$$= \left|\frac{r^3\cos^2\theta(\cos\theta - \sin\theta)}{r^2(1 + \cos\theta\sin\theta)}\right| = r\cos^2\theta\frac{|\cos\theta - \sin\theta|}{|1 + \sin\theta\cos\theta|}.$$

It is actually sufficient to stop here: The denominator is not zero and the numerator is bounded and proportional to r, which converges to zero. Using basic trigonometry results,

$$\left|\frac{x^3 - x^2y}{x^2 + y^2 + xy} - 0\right| = r\cos^2\theta\frac{\sqrt{2}|\cos\theta\cos(\pi/4) - \sin\theta\sin(\pi/4)|}{\left|1 + \frac{1}{2}2\sin\theta\cos\theta\right|}$$
$$\leq r\sqrt{2}\frac{|\cos(\theta + \pi/4)|}{\left|1 + \frac{1}{2}\sin(2\theta)\right|} \leq r\frac{\sqrt{2}}{1/2} = 2\sqrt{2}r \to 0 \text{ as } r \to 0.$$

Thus, given $\epsilon > 0$, however, small, we can find a δ, as a function of ϵ, e.g., $\delta = \epsilon/(2\sqrt{2})$, such that

$$\left|\frac{x^3 - x^2y}{x^2 + y^2 + xy} - 0\right| < \epsilon \quad \text{whenever} \quad r < \delta.$$

Given that we have invoked an arbitrary curve whose sole requirement is to pass through the limit point (the origin) the result is general, the limit exists, and is indeed 0. ■

The following results parallel their univariate counterparts, and we state them without proof. Unless otherwise specified, let $\mathbf{f}, \mathbf{g} : A \subset \mathbb{R}^n \to \mathbb{R}^m$.

- Assume for $\mathbf{x}_0 \in A$ that $\lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{f}(\mathbf{x})$ and $\lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{g}(\mathbf{x})$ exist (which means, exists in $\mathbb{R}^m$). Then, for constant values $k_1, k_2 \in \mathbb{R}$, limits satisfy linearity and homogeneity, i.e., mixing the two,

$$\lim_{\mathbf{x}\to\mathbf{x}_0} (k_1\mathbf{f} + k_2\mathbf{g})(\mathbf{x}) = k_1 \lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{f}(\mathbf{x}) + k_2 \lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{g}(\mathbf{x}), \tag{4.14}$$

  which means functions with limits at $\mathbf{x}_0 \in A$ form a vector space.

- Analogous to limits of products of functions, each of which has a limit, we have

$$\lim_{\mathbf{x}\to\mathbf{x}_0} (\mathbf{f} \cdot \mathbf{g})(\mathbf{x}) = \lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{f}(\mathbf{x}) \cdot \lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{g}(\mathbf{x}), \tag{4.15}$$

  where, for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^n x_i y_i$ is the dot product of $\mathbf{x}$ and $\mathbf{y}$, as in (3.1).

- Let $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$. If $\mathbf{f} : A \to \mathbb{R}^m$ and $\mathbf{g} : B \to \mathbb{R}^p$ such that $\mathbf{f}(A) \subset B$, then the composite function $\mathbf{g} \circ \mathbf{f}$ is well-defined. If $\mathbf{y}_0 := \lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{f}(\mathbf{x})$ and $\lim_{\mathbf{y}\to\mathbf{y}_0} \mathbf{g}(\mathbf{y})$ both exist, then $\lim_{\mathbf{x}\to\mathbf{x}_0} (\mathbf{g} \circ \mathbf{f})(\mathbf{x}) = \lim_{\mathbf{y}\to\mathbf{y}_0} \mathbf{g}(\mathbf{y})$.

- Let $\mathbf{f} : A \to \mathbb{R}^m$ with $A \subset \mathbb{R}^n$. Paralleling the univariate case (2.19), mapping $\mathbf{f}$ is *continuous* at $\mathbf{a} \in A$ if

$$\lim_{\mathbf{x}\to\mathbf{a}} \mathbf{f}(\mathbf{x}) = \mathbf{f}\left(\lim_{\mathbf{x}\to\mathbf{a}} \mathbf{x}\right) = \mathbf{f}(\mathbf{a}). \tag{4.16}$$

  Equivalently, and as proven in the univariate case in §2.1, $\mathbf{f}$ is continuous at $\mathbf{a} \in A$ if, for a given $\epsilon > 0$, $\exists \delta > 0$ such that, if $\|\mathbf{x} - \mathbf{a}\| < \delta$ and $\mathbf{x} \in A$, then $\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{a})\| < \epsilon$. We state this important result yet more explicitly: From Fitzpatrick (2009, Thm 11.11), who uses the notation $\text{dist}(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|$:

  > Let A be a subset of $\mathbb{R}^n$ that contains the point $\mathbf{u}$. Then the following two assertions about a mapping $F : A \to \mathbb{R}^m$ are equivalent:
  >
  > i. The mapping $F : A \to \mathbb{R}^m$ is continuous at the point $\mathbf{u}$; that is, for a sequence $\{\mathbf{u}_k\}$ in A,
  >
  > $$\lim_{k\to\infty} \text{dist}(\mathbf{u}_k, \mathbf{u}) = 0 \implies \lim_{k\to\infty} \text{dist}(F(\mathbf{u}_k), F(\mathbf{u})) = 0. \tag{4.17}$$
  >
  > ii. For each positive number ϵ there is a positive number δ such that, for a point $\mathbf{v}$ in A,
  >
  > $$\text{dist}(\mathbf{v}, \mathbf{u}) < \delta \implies \text{dist}(F(\mathbf{v}), F(\mathbf{u})) < \epsilon. \tag{4.18}$$

- If $\mathbf{f}$ is continuous at every point in its domain A, then $\mathbf{f}$ is said to be continuous, and we write $\mathbf{f} \in \mathcal{C}^0$ or, more accurately, $\mathbf{f} \in \mathcal{C}^0(A)$.

- Mapping $\mathbf{f}$ is *uniformly continuous* on subset $S \subset A$ if: for a given $\epsilon > 0$, $\exists \delta > 0$ such that, if $\mathbf{x}, \mathbf{y} \in S$, and $\|\mathbf{x} - \mathbf{y}\| < \delta$, then $\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\| < \epsilon$.

- Let $\mathbf{f}, \mathbf{g} : A \subset \mathbb{R}^n \to \mathbb{R}^m$, and $k_1, k_2 \in \mathbb{R}$. Similar to the above results (4.14) and (4.15) on limits, if $\mathbf{f}$ and $\mathbf{g}$ are continuous at $\mathbf{x}_0 \in A$, then so are $k_1\mathbf{f} + k_2\mathbf{g}$ and $\mathbf{f} \cdot \mathbf{g}$ at $\mathbf{x}_0$. If $\mathbf{f}$ and $\mathbf{g}$ are continuous (meaning, as stated above, continuous at all points in their domain), then $k_1\mathbf{f} + k_2\mathbf{g}$ is continuous. This means that continuous functions (with the same domain and range) form a vector space.

- Let $f : A \subset \mathbb{R}^n \to \mathbb{R}$. We say that f has a relative maximum at $\mathbf{a} \in A$ if there exists an n-ball $U \subset A$ such that $f(\mathbf{a}) \geq f(\mathbf{x})$ for all $\mathbf{x} \in U$. The definition for a relative minimum is analogous.

- Similar to (2.29) in the univariate case, if $A \subset \mathbb{R}^n$ is closed and bounded,[28] and $f : A \to \mathbb{R}$ is continuous, then f takes on minimum and maximum values. That is,

$$\exists \mathbf{a}, \mathbf{b} \in A \textbf{ such that, } \forall \mathbf{x} \in A, \; f(\mathbf{a}) \leq f(\mathbf{x}) \leq f(\mathbf{b}). \tag{4.19}$$

- The intermediate value theorem (IVT) given in (2.30) for univariate functions can be generalized in an immediate way as follows. Let $f : A \to \mathbb{R}$ be continuous on subset $S \subset A \subset \mathbb{R}^n$, where S is a closed box in $\mathbb{R}^n$, e.g., the set $[a, b] \times [c, d] \subset \mathbb{R}^2$. Notice that, for $n = 1$, S is a closed, bounded interval, as was required in the IVT for univariate functions. Assume $\mathbf{a}, \mathbf{b} \in S$. Let $\alpha = f(\mathbf{a})$ and $\beta = f(\mathbf{b})$. Given a number γ with $\alpha < \gamma < \beta$, $\exists \mathbf{c} \in S$ such that $f(\mathbf{c}) = \gamma$.

  However, for $n > 1$, other specifications for S are possible such that the IVT holds. Characterizing the IVT in this case entails introducing the concepts of connectedness, polygonally connected, and regions. See, e.g., Trench (2012, pp. 295-7) and, particularly, Petrovic, Advanced Calculus: Theory and Practice, 2nd ed., 2020, §10.6, for clear, detailed presentations.

  Recall that continuity alone is not sufficient for the IVT for $n = 1$. As an example, let domain $D = \mathbb{R} \setminus \{0\}$, and $f : D \to \mathbb{R}$ be given by $f(x) = -1$ for $x < 0$ and $f(x) = 1$ for $x > 0$. Function f is continuous on D, but there is no x_0 such that $f(x_0)$ takes on a value strictly between -1 and 1. The next example below illustrates the same idea for $n = 2$.

  A set $A \subset \mathbb{R}^n$ is said to be *polygonally connected* if, for any two points $P, Q \in A$ there exists a polygonal line that connects them. More precisely, there exists a positive integer n and points $P_0 = P, P_1, \ldots, P_n = Q$ all in A, such that each line segment $P_iP_{i+1}, 0 \leq i \leq n-1$, completely lies within A. Note that a closed box in $\mathbb{R}^n$ is polygonally connected. The set $A = \{(x, y) \in \mathbb{R}^2 : 1 \leq x^2 + y^2 \leq 4\}$ is polygonally connected.

  The n-dimensional IVT, as stated and proved in Petrovic, p. 327, is: Let f be a continuous function on a polygonally connected domain $A \subset \mathbb{R}^n$, and let $P, Q \in A$. If $f(P) < 0$ and $f(Q) > 0$ then there exists $M \in A$ such that $f(M) = 0$.

[28] According to the Heine-Borel theorem, this is equivalent to set A being compact. Proofs of this and related results for functions on compact sets are best done invoking the machinery for sequential and topological compactness. This is the reason we do not give proofs here, but they are available in (nearly) all textbooks on multivariate analysis.

Example 4.7 *(Petrovic, p. 324) (A continuous function that does not have the IVP) Prove that the function*

$$f(x,y) = \begin{cases} -1, & \text{if } x^2+y^2 < 1, \\ 2, & \text{if } x^2+y^2 > 4, \end{cases}$$

is continuous, but there is no point (c_1, c_2) *such that* $f(c_1, c_2) = 0$.

Solution. The domain of f *consists of the open unit disk and the outside of the disk centered at the origin and of radius 2. It is obvious that* f *is continuous at every point of its domain. Further,* $f(0,0) = -1$ *and* $f(2,3) = 2$ *(because* $2^2 + 3^2 = 13 > 4$*). Yet, there is no point* (c_1, c_2) *such that* $f(c_1, c_2) = 0$. ■

Again paralleling the univariate case, we need to address continuity of the composition of (two) functions. First recall the image of a subset: Given a mapping $F : A \to \mathbb{R}^m$, if B is a subset of the domain A, the image of the set B under the mapping $F : A \to \mathbb{R}^m$, denoted by $F(B)$, is defined by the formula

$$F(B) \equiv \{\mathbf{v} \text{ in } \mathbb{R}^m \mid \mathbf{v} = F(\mathbf{u}) \quad \text{for some } \mathbf{u} \text{ in } B\}.$$

As in Fitzpatrick (2009, Thm 11.5),

> Let A be a subset of $\mathbb{R}^n$ that contains the point $\mathbf{u}$. Suppose that the mapping $G : A \to \mathbb{R}^m$ is continuous at the point $\mathbf{u}$. Let B be a subset of $\mathbb{R}^m$ with $G(A) \subseteq B$ and suppose that the mapping $H : B \to \mathbb{R}^k$ is continuous at the point $G(\mathbf{u})$. Then the composition $H \circ G : A \to \mathbb{R}^k$ is continuous at $\mathbf{u}$.

> Proof: Let $\{\mathbf{u}_k\}$ be a sequence in A that converges to the point $\mathbf{u}$. Since the mapping $G : A \to \mathbb{R}^m$ is continuous at $\mathbf{u}$, it follows that the image sequence $\{G(\mathbf{u}_k)\}$ converges to $G(\mathbf{u})$. But then $\{G(\mathbf{u}_k)\}$ is a sequence in B that converges to the point $G(\mathbf{u})$. The continuity of the mapping $H : B \to \mathbb{R}^k$ at the point $G(\mathbf{u})$ implies that the sequence $\{H(G(\mathbf{u}_k))\}$ converges to $H(G(\mathbf{u}))$; that is, the sequence $\{(H \circ G)(\mathbf{u}_k)\}$ converges to $(H \circ G)(\mathbf{u})$.

Example 4.8 *Define the function* $f : \mathbb{R}^n \to \mathbb{R}$ *by* $f(\mathbf{u}) = \|\mathbf{u}\|$*, for* $\mathbf{u} \in \mathbb{R}^n$*. Then the function* $f : \mathbb{R}^n \to \mathbb{R}$ *is continuous. To see this, recall the definition of projection functions in (4.5), and observe that, with* $p_i(\mathbf{u}) = u_i$*,* $i = 1, \ldots, n$*,*

$$f = h \circ (p_1 p_1 + \cdots + p_n p_n) : \mathbb{R}^n \to \mathbb{R},$$

where $h(x) = \sqrt{x}$ *for* $x \geq 0$*. Since products, sums, and compositions of continuous maps are again continuous, it follows that the function* $f : \mathbb{R}^n \to \mathbb{R}$ *is continuous. Thus, for* $\mathbf{x}_*$ *a point in* $\mathbb{R}^n$*,* $\lim_{\mathbf{x} \to \mathbf{x}_*} \|\mathbf{x}\| = \|\mathbf{x}_*\|$. ■

We state the following "Structure of Open Sets" results, as they are fundamental in analysis, but we will not make use of them:

Theorem: If U is an open subset of $\mathbb{R}$, then there exists a finite or countable collection $\{I_j\}$ of pairwise disjoint open intervals such that $U = \bigcup_n I_j$. See, e.g., Terrell Theorem 4.1.6; or Stoll, Thm 2.2.20, for proof. This is extended to $\mathbb{R}^n$ as follows, from Terrell, p. 519:

Theorem (Structure of Open Sets in $\mathbb{R}^n$): Every open set in $\mathbb{R}^n, n \geq 1$, can be expressed as a countable union of nonoverlapping closed cubes.

4.2 Partial Derivatives and the Gradient

> In this section we deal with functions of one variable. The multivariable case in which $f : \mathbb{R}^n \to \mathbb{R}$ offers no new ideas, only new notation.
>
> (Charles Pugh, Real Mathematical Analysis, 2nd edition, 2015, p. 406)

This opening quote from Pugh, in his §6.6, regards the Lebesgue integral, and indeed, one of the luxuries of that integral is that the multivariate case is very easy to handle, once the framework in the univariate case is established. The reason I include this quote here is that, when it comes to differentiation for multivariate functions, it is *not* the case that "only new notation" is required; and in fact, quite some new ideas and concepts emerge. We will see some new ideas in this section, but notably in subsequent sections, such as for directional derivatives and the multivariate Mean Value Theorem.

Let $f : A \to \mathbb{R}$ with $A \subset \mathbb{R}^n$ an open set. For every $\mathbf{x} = (x_1, \ldots, x_n) \in A$ and for each $i = 1, 2, \ldots, n$, the *partial derivative* of f with respect to x_i is defined as

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) = \lim_{h \to 0} \frac{f(x_1, \ldots, x_{i-1}, x_i + h, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_n)}{h}, \tag{4.20}$$

if the limit exists. Because the remaining $n - 1$ variables in $\mathbf{x}$ are held constant, the partial derivative is conceptually identical to the Newton quotient (2.32) for univariate functions. This can be more compactly written by defining $\mathbf{e}_i = (0, 0, \ldots, 0, 1, 0, \ldots, 0)$ to be the n-length vector with a one in the ith position, and zero elsewhere, so that

$$(D_i f)(\mathbf{x}) := \frac{\partial f}{\partial x_i}(\mathbf{x}) = \lim_{h \to 0} \frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h}. \tag{4.21}$$

As indicated, a popular and useful alternative notation for the partial derivative is $D_i f(\mathbf{x})$ or, better, $(D_i f)(\mathbf{x})$, with the advantage that the name of the ith variable (in this case x_i) does not need to be explicitly mentioned. This is termed the partial derivative of f with respect to the ith variable, at $\mathbf{x}$.

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$. Then the function $f : \mathcal{O} \to \mathbb{R}$ is said to have *first-order partial derivatives* provided that, for each index i with $1 \leq i \leq n$, the function $f : \mathcal{O} \to \mathbb{R}$ has a partial derivative with respect to its ith component, at every point in $\mathcal{O}$.

If each of the n partial derivatives exists at $\mathbf{x}$, then the *gradient* of f at $\mathbf{x}$, denoted $(\operatorname{grad} f)(\mathbf{x})$ (and rhyming with sad and glad), or $(\nabla f)(\mathbf{x})$, is the row vector of all partial derivatives:

$$(\nabla f)(\mathbf{x}) = (\operatorname{grad} f)(\mathbf{x}) := (D_1 f(\mathbf{x}), \ldots, D_n f(\mathbf{x}))\,. \tag{4.22}$$

Some further insight into partial derivatives is gained by using a *parametrized path*, allowing for a consideration that parallels the univariate case. In particular, recall from (2.32) and (2.34) that, for $f : D \subset \mathbb{R} \to \mathbb{R}$ differentiable on D^o (the interior of D), and a point $x \in D^o$, the derivative of f at x is the Newton quotient

$$f'(x) = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h} = \lim_{y \to x} \frac{f(y) - f(x)}{y - x}. \tag{4.23}$$

Below, we produce equation (4.24) which is a direct analog of this.

The following presentation is based on Fitzpatrick (2009, §13.2). It is of use to review the concept of line segment in $\mathbb{R}^n$ from Example 4.1. Own notes are in blue color.

Consider a real-valued function of several real variables $f : \mathbb{R}^n \to \mathbb{R}$, together with two points $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$. Suppose that we want to compare $f(\mathbf{u})$ with $f(\mathbf{v})$. When $n = 1$ and the function $f : \mathbb{R} \to \mathbb{R}$ is differentiable, we can use the Mean Value Theorem to compare these two values. When $n > 1$, the following restriction procedure is natural. Look at the parametrized segment from $\mathbf{u}$ to $\mathbf{v}$—that is, the parametrized path $\gamma : [0, 1] \to \mathbb{R}^n$ defined by

$$\gamma(t) = \mathbf{u} + t(\mathbf{v} - \mathbf{u}) = t\mathbf{v} + (1 - t)\mathbf{u}, \quad \text{for} \quad 0 \leq t \leq 1.$$

Then consider the composition of the function $f : \mathbb{R}^n \to \mathbb{R}$ with this parametrized path, which is the function $\psi : [0, 1] \to \mathbb{R}$ defined by

$$\psi(t) = f(\mathbf{u} + t(\mathbf{v} - \mathbf{u})), \quad \text{for} \quad 0 \leq t \leq 1,$$

with $\psi(0) = f(\mathbf{u})$ and $\psi(1) = f(\mathbf{v})$. Thus, to compare $f(\mathbf{u})$ with $f(\mathbf{v})$ is to compare $\psi(0)$ with $\psi(1)$. If we can determine that $\psi : [0, 1] \to \mathbb{R}$ is continuous and that $\psi : (0, 1) \to \mathbb{R}$ is differentiable, then we can apply the Mean Value Theorem for functions of a single variable to compare $f(\mathbf{u})$ with $f(\mathbf{v})$. Thus, it is necessary to investigate the properties of the function $f : \mathbb{R}^n \to \mathbb{R}$ that will allow us to conclude that the above auxiliary function $\psi : [0, 1] \to \mathbb{R}$ is continuous; to conclude that $\psi : (0, 1) \to \mathbb{R}$ is differentiable; and to compute $\psi' : (0, 1) \to \mathbb{R}$.

We can regard $\psi : [0, 1] \to \mathbb{R}$ as being the restriction of the function $f : \mathbb{R}^n \to \mathbb{R}$ to the line segment between the points $\mathbf{u}$ and $\mathbf{v}$, together with the placing of a coordinate system on this line segment. In the case where $n = 2$, the graph of $\psi : [0, 1] \to \mathbb{R}$ is obtained by intersecting the graph of $f : \mathbb{R}^2 \to \mathbb{R}$ with the plane that is parallel to the z-axis and contains the segment joining $\mathbf{u}$ and $\mathbf{v}$. For this reason, we refer to the function $\psi : [0, 1] \to \mathbb{R}$ as a *section of the function* $f : \mathbb{R}^n \to \mathbb{R}$.

In order to analyze the differentiability of the function $\psi : (0, 1) \to \mathbb{R}$ at the point t_0, we change variables by setting $\mathbf{x} = \mathbf{u} + t_0(\mathbf{v} - \mathbf{u})$, $\mathbf{p} = \mathbf{v} - \mathbf{u}$, and $s = t - t_0$; then

$$\frac{\psi(t) - \psi(t_0)}{t - t_0} = \frac{f(\mathbf{x} + s\mathbf{p}) - f(\mathbf{x})}{s},$$

and taking limits and noting $s \to 0$ is the same as $t \to t_0$, we have the analog of (4.23),

$$\psi'(t_0) = \lim_{t \to t_0} \frac{\psi(t) - \psi(t_0)}{t - t_0} = \lim_{s \to 0} \frac{f(\mathbf{x} + s\mathbf{p}) - f(\mathbf{x})}{s}, \tag{4.24}$$

provided that the limit exists. The strategy of looking at sections of a function, together with (4.24), motivates the introduction of the following concept of a partial derivative.

Definition: Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}$, and let i be an index with $1 \leq i \leq n$. A function $f : \mathcal{O} \to \mathbb{R}$ is said to *have a partial derivative with respect to its ith component at the point* $\mathbf{x}$ provided that the limit

$$\lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{e}_i) - f(\mathbf{x})}{t} \tag{4.25}$$

exists, where, instead of generic point $\mathbf{p}$ in (4.24), $\mathbf{e}_i$ is the ith unit vector, i.e., the vector whose ith component is 1 and whose other components are 0. If this limit exists, then we denote its value by $\partial f / \partial x_i(\mathbf{x})$, and call it the *partial derivative of $f : \mathcal{O} \to \mathbb{R}$ with respect to the ith component, at the point* $\mathbf{x}$.

The geometric meaning of $\partial f/\partial x_i(\mathbf{x})$ is as follows: Choose a positive number r such that the open ball $\mathcal{B}_r(\mathbf{x})$ is contained in $\mathcal{O}$ and consider the section defined by $\psi(t;i,\mathbf{x}) = \psi(t) = f(\mathbf{x}+t\mathbf{e}_i)$, for $|t|<r$. Then $f:\mathcal{O}\to\mathbb{R}$ has a partial derivative with respect to its ith component at the point $\mathbf{x}$ precisely when there is a tangent line to the graph of this section at the point on the graph corresponding to $t=0$, at which point the slope of this tangent is the number

$$\psi'(0) = \frac{\partial f}{\partial x_i}(\mathbf{x}),$$

or, in more detail, and of value for directional derivatives below,

$$\psi'(0) = \frac{d}{dt}\psi(t)\bigg|_{t=0} = \frac{d}{dt}f(\mathbf{x}+t\mathbf{e}_i)\bigg|_{t=0} = \frac{\partial f}{\partial x_i}(\mathbf{x}) = (D_i f)(\mathbf{x}). \tag{4.26}$$

Thus, the existence of $\partial f/\partial x_i(\mathbf{x})$ is equivalent to the differentiability of a function of a single real variable, so we can immediately use the single-variable differentiation results to obtain addition, product, and quotient rules for partial derivatives.

Recall from (2.42) that, in the univariate case, if f is differentiable at a, then f is continuous at a; and if f is differentiable on its entire domain D, then f is continuous on D. This is not necessarily the case for $n>1$: A function $f:D\subset\mathbb{R}^n\to\mathbb{R}$ that has first-order partial derivatives at all points in the (interior of the) domain need not be continuous. The next example gives a case in point. The key is that, while $\forall\mathbf{x}\in D$, $(\operatorname{grad} f)(\mathbf{x})$ exists, it is not continuous on all of D.

Example 4.9 *Recall the bivariate function from Example 4.3,*

$$f(x,y) = \begin{cases} xy/\left(x^2+y^2\right), & \text{if } (x,y)\neq(0,0),\\ 0, & \text{if } (x,y)=(0,0),\end{cases}$$

in which we showed that $\lim_{\mathbf{x}\to\mathbf{0}} f(\mathbf{x})$ *does not exist. Repeating from Example 4.3, observe that the sequence* $\{(1/k,1/k)\}$ *converges to* $(0,0)$ *and that* $f(1/k,1/k)=1/2$ *for each index* k*, so the image sequence* $\{f(1/k,1/k)\}$ *converges to* $1/2$*. But* $1/2\neq f(0,0)$*. Thus, the function* f *is not continuous at the point* $(0,0)$*.*

Using (4.25) *with* $(x,y)=(0,0)$*,*

$$\frac{\partial f}{\partial x}(0,0) = \lim_{h\to0}\frac{f(h,0)-f(0,0)}{h} = \lim_{h\to0}\frac{0-0}{h} = 0;$$

and, from symmetry, the same is true for $(D_2f)(0,0)$*. Thus,* $(\operatorname{grad} f)(0,0)$ *exists, and equals* $(0,0)$*.*

For $(x,y)\neq(0,0)$*, there is a neighborhood of* (x,y) *on which the restriction of* $f:\mathbb{R}^2\to\mathbb{R}$ *is a quotient of polynomials whose denominator does not vanish. Thus, on that neighborhood,* $\partial f/\partial x(x,y)$ *and* $\partial f/\partial y(x,y)$ *exist; moreover, a short computation yields*

$$\frac{\partial f}{\partial x}(x,y) = \frac{y^3-x^2y}{\left(x^2+y^2\right)^2} \quad \text{and} \quad \frac{\partial f}{\partial y}(x,y) = \frac{x^3-y^2x}{\left(x^2+y^2\right)^2}.$$

Thus, the function f *has first-order partial derivatives at every point in the plane* $\mathbb{R}^2$*. However, these two derivatives are not continuous at* $(0,0)$*. To see this, compare their limiting behavior for sequences* $(0,1/k)$ *and* $(1/k,0)$*.*

Below, we will see that this lack of continuity in the partial derivatives means we cannot conclude that f *is continuous at all points in its domain; and indeed, this is true for this function* f*.* ■

4.3 Differentiability and Tangent Maps

> One thing you will observe about all these books – they use pictures to convey the mathematical ideas. Beware of books that don't.
>
> Charles Pugh, Real Mathematical Analysis, 2nd edition, 2015, p. 467

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ and let the function $f : \mathcal{O} \to \mathbb{R}$ be such that it has first-order partial derivatives. If in addition, grad f (i.e., each of $(D_i f)(\mathbf{x}) : \mathcal{O} \to \mathbb{R}$, $1 \leq i \leq n$) is continuous, then we will see below that f is continuous. Since this additional assumption will play an important part later, it is useful to name it:

Definition: Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$. Then a function $f : \mathcal{O} \to \mathbb{R}$ is said to be *continuously differentiable*, provided that it has first-order partial derivatives for all $\mathbf{x} \in \mathcal{O}$, and such that each partial derivative $(D_i f)(\mathbf{x}) : \mathcal{O} \to \mathbb{R}$ is continuous, $1 \leq i \leq n$.

We also state now the fundamental result on continuity, albeit whose proof needs to wait until other results are proven. This result is shown below in (4.46), and then, with a different, more sophisticated proof, using the multivariate Mean Value Theorem in §4.5, on page 229.

Theorem (Continuity): Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ and suppose that the function $f : \mathcal{O} \to \mathbb{R}$ is continuously differentiable. Then the function $f : \mathcal{O} \to \mathbb{R}$ is continuous.

Example 4.10 *Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ with $f(0,0) = 0$ and*

$$f(x,y) = \frac{x^2 - y^2}{x^2 + y^2}, \quad (x,y) \neq (0,0).$$

To assess continuity, we compute

$$\lim_{x\to 0} f(x,0) = \lim_{x\to 0} \frac{x^2}{x^2} = \lim_{x\to 0} 1 = 1, \qquad \lim_{y\to 0} f(0,y) = \lim_{y\to 0} \frac{-y^2}{y^2} = \lim_{y\to 0} -1 = -1,$$

showing that f is not continuous at $(0,0)$, and there is no way to redefine its value at $(0,0)$ to make it continuous. From the contrapositive of the continuity theorem, we know that at least one of the first order partial derivatives is not continuous on the whole domain. ■

Example 4.11 *Now consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ with $f(0,0) = 0$ and*

$$f(x,y) = xy\frac{x^2 - y^2}{x^2 + y^2}, \quad (x,y) \neq (0,0).$$

We compute

$$\lim_{x\to 0} f(x,0) = \lim_{x\to 0} 0 = 0, \qquad \lim_{y\to 0} f(0,y) = \lim_{y\to 0} 0 = 0,$$

i.e., "so far so good", and continue to check: Letting $y = ax$ for some $a \in \mathbb{R}$,

$$\lim_{x\to 0} f(x,ax) = ax^2 \frac{x^2(1-a^2)}{x^2(1+a^2)} = a\frac{(1-a^2)}{(1+a^2)} \lim_{x\to 0} x^2 = 0,$$

so that the limit of the function on every linear path to $(0,0)$ is zero. Further, nonlinear paths would need to be inspected, but we omit these, and presume that f is indeed continuous at $(0,0)$. With f being a ratio of polynomials for $(x,y) \neq (0,0)$, it is thus continous at all points in its domain $\mathbb{R}^2$. The continuity theorem is only an "if then", so that we cannot

conclude that the first order partial derivatives are continous. Instead, we have to check. Straightforward calculation shows that

$$\frac{\partial f}{\partial x} = \frac{yx^4 - y^5 + 4x^2y^3}{\left(x^2+y^2\right)^2}, \qquad \frac{\partial f}{\partial y} = \frac{x^5 - xy^4 - 4x^3y^2}{\left(x^2+y^2\right)^2}.$$

To assess their continuity, note first that

$$\lim_{x\to 0}\frac{\partial f}{\partial x}(x,0) = \lim_{x\to 0}\frac{0}{x^4} = \lim_{x\to 0} 0 = 0, \qquad \lim_{y\to 0}\frac{\partial f}{\partial x}(0,y) = \lim_{y\to 0} -y = 0,$$

so that, if the limit of $(\partial f/\partial x)(x,y)$ *as* $(x,y)\to(0,0)$ *exists, it must be zero. Let* $y = ax$ *for some* $a\in\mathbb{R}$*, so that*

$$\lim_{x\to 0}\frac{\partial f}{\partial x}(x,ax) = \lim_{x\to 0}\frac{ax^5 - a^5x^5 + 4a^3x^5}{x^4\left(1+a^2\right)^2} = \frac{a\left(4a^2 - a^4 + 1\right)}{\left(1+a^2\right)^2}\lim_{x\to 0} x = 0,$$

confirming that, at least along all lines approaching $(0,0)$*, the limit of* $\partial f/\partial x$ *is zero. One could and should check other, nonlinear, paths to be sure. Next, observe that*

$$\frac{\partial f}{\partial x}(0,0) = \lim_{h\to 0}\frac{f(h,0) - f(0,0)}{h} = \lim_{h\to 0}\frac{0-0}{h} = 0.$$

Thus, it is presumable (we did not check all paths) that $\partial f/\partial x$ *is continuous at zero, and thus, being a ratio of polynomials for* $(x,y)\neq(0,0)$*, is continuous everywhere on the domain* $\mathbb{R}^2$*.*

Similarly for $\partial f/\partial y$*,*

$$\lim_{x\to 0}\frac{\partial f}{\partial y}(x,0) = \lim_{x\to 0}\frac{x^5}{x^4} = 0, \qquad \lim_{y\to 0}\frac{\partial f}{\partial y}(0,y) = \lim_{y\to 0}\frac{0}{y^4} = \lim_{y\to 0} 0 = 0,$$

and, $y = ax$ *for some* $a\in\mathbb{R}$*,*

$$\lim_{x\to 0}\frac{\partial f}{\partial y}(x,ax) = \lim_{x\to 0}\frac{x^5 - a^4x^5 - 4x^5a^2}{x^4\left(1+a^2\right)^2} = \frac{\left(1 - a^4 - 4a^2\right)}{\left(1+a^2\right)^2}\lim_{x\to 0} x = 0.$$

Also,

$$\frac{\partial f}{\partial y}(0,0) = \lim_{h\to 0}\frac{f(0,h) - f(0,0)}{h} = \lim_{h\to 0}\frac{0-0}{h} = 0.$$

Assuming the adequacy of this inspection that grad f *is continuous, the above theorem implies that* f *is continuous, which we confirmed above.* ■

In the following, let $f : A\to\mathbb{R}$ with $A\subset\mathbb{R}^n$ an open set and such that $(\operatorname{grad} f)(\mathbf{x})$ exists $\forall\mathbf{x}\in A$. Recall from (2.33) and (2.287) that, for $n = 1$, the *tangent to the curve* at the point (x_0,y_0), for $x_0\in A\subset\mathbb{R}$ and $y_0 = f(x_0)$ is the (non-vertical) line $T(x) = y_0 + f'(x_0)(x - x_0)$. This is the best linear approximation to f in a neighborhood of x_0 such that $f(x_0) = T(x_0)$, and, from (2.32) and (2.33), satisfies

$$\lim_{x\to x_0}\frac{f(x) - T(x)}{x - x_0} = 0.$$

For $n = 2$, envisioning a thin, flat board resting against a sphere in 3-space, we seek a (non-vertical) plane in $\mathbb{R}^3$ that is "tangent" to f at a given point, say (x_0, y_0, z_0), for $(x_0, y_0) \in A$ and $z_0 = f(x_0, y_0)$. A plane is linear in both x and y, so its equation is, recalling (3.8),

$$z = z_0 + s(x - x_0) + t(y - y_0),$$

where s and t need to be determined.

When restricted to the plane $y = y_0$, the surface f is just the curve $z = g(x) := f(x, y_0)$ in $\mathbb{R}^2$, and the plane we seek is just the line $z = z_0 + s(x - x_0)$. This is the $n = 1$ case previously discussed, so the tangent to the curve $g(x)$ at x_0 is the line $z = z_0 + g'(x_0)(x - x_0)$, i.e., $s = D_1 f(x_0, y_0)$. Similarly, $t = D_2 f(x_0, y_0)$. This gives rise to the definition, in the $n = 2$ case, of the *tangent plane* of f at (x_0, y_0, z_0):

Definition: For function $f : D \subset \mathbb{R}^2 \to \mathbb{R}$ and $(x_0, y_0) \in D^o$, the tangent plane at (x_0, y_0, z_0), where $z_0 = f(x_0, y_0)$, is the linear function

$$T(x, y) = f(x_0, y_0) + (D_1 f(x_0, y_0))(x - x_0) + (D_2 f(x_0, y_0))(y - y_0) \tag{4.27}$$

that satisfies

$$f(x_0, y_0) = T(x_0, y_0), \quad \text{and} \quad \lim_{(x,y)\to(x_0,y_0)} \frac{f(x, y) - T(x, y)}{\|(x, y) - (x_0, y_0)\|} = 0. \tag{4.28}$$

The reason for (the more strenuous condition of) dividing $f(x, y) - T(x, y)$ in (4.28) by

$$\|(x, y) - (x_0, y_0)\| = \sqrt{(x - x_0)^2 + (y - y_0)^2}$$

is discussed below.

Let $\mathbf{v}_1$ be the vector in the xz plane, on the slice in 3D space $y = y_0$, originating from the origin, with endpoint $(1, 1 \times D_1 f(x_0, y_0))$, and thus parallel to the tangent line L_1 given by $z = z_0 + (D_1 f)(x_0, y_0)(x - x_0)$. Vector $\mathbf{v}_2$ in the yz plane, and line L_2, are similarly defined. Motivated by the quote at the beginning of this subsection regarding the importance of pictures for conveying mathematical ideas, we illustrate these lines and vectors in Figures 32 and 33, taken from Miklavcic's excellent and accessible *An Illustrative Guide to Multivariable and Vector Calculus* (2020).

Notice that $\mathbf{v}_1$ and $\mathbf{v}_2$ cannot be parallel to each other. The vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ define a tangent plane, T. Capitalizing on our discussion of the cross product, from (3.28) or (3.33), this tangent plane, for $n = 2$, has normal vector

$$\mathbf{n} = \mathbf{v}_1 \times \mathbf{v}_2 = \begin{vmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \\ 1 & 0 & \frac{\partial f}{\partial x}\Big|_{(x_0,y_0)} \\ 0 & 1 & \frac{\partial f}{\partial y}\Big|_{(x_0,y_0)} \end{vmatrix} = -\frac{\partial f}{\partial x}\Big|_{(x_0,y_0)} \mathbf{e}_1 - \frac{\partial f}{\partial y}\Big|_{(x_0,y_0)} \mathbf{e}_2 + \mathbf{e}_3. \tag{4.29}$$

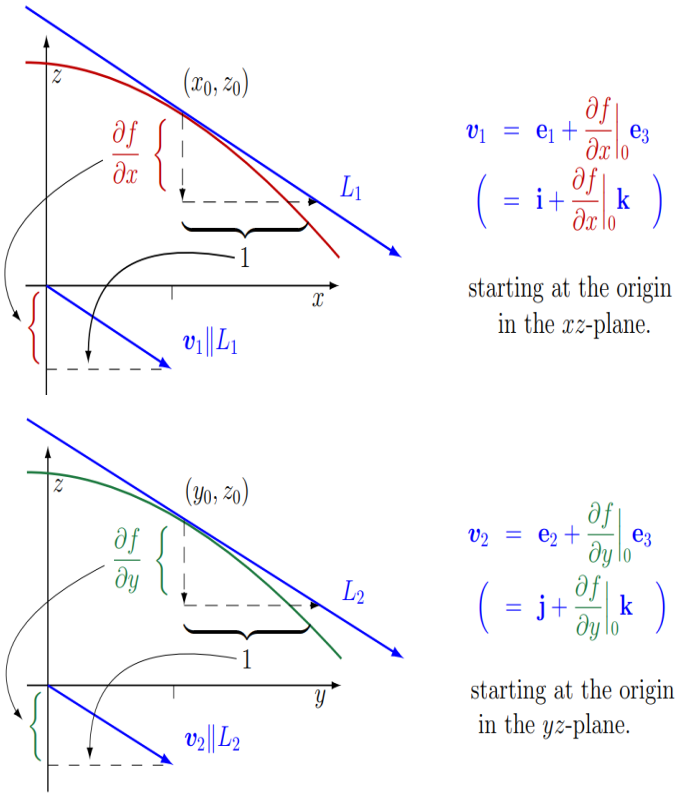


Figure 32: Top: A tangent vector and line in the x-direction. In the cross section parallel to the xz-plane, a vector parallel to line L_1 is $\mathbf{v}_1$. Bottom: A tangent vector and line in the y-direction. In the cross section parallel to the yz-plane, a vector parallel to line L_2 is $\mathbf{v}_2$. Taken from Miklavcic, p. 65.

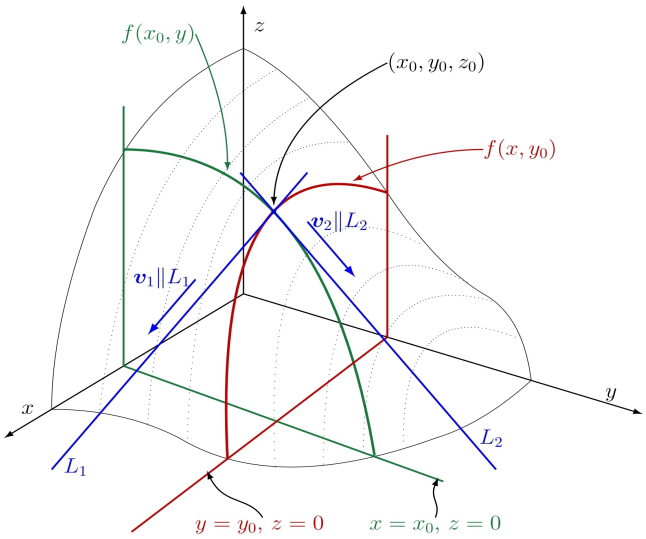


Figure 33: The two components of the tangent plane for a function from $\mathbb{R}^2$ to $\mathbb{R}$. Taken from Miklavcic, p. 66

The plane defined by L_1 and L_2

1. is tangent to the surface $z = f$ at (x_0, y_0, z_0);
2. is spanned by $\mathbf{v}_1$ and $\mathbf{v}_2$; and
3. has the same normal as the normal to the graph of $z = f(x, y, z)$ at $\mathbf{x}_0 = (x_0, y_0, z_0)$.

Its equation can be found from the scalar vector product $\mathbf{n} \cdot (\mathbf{x} - \mathbf{x}_0) = 0$, i.e.,

$$z - z_0 = \left.\frac{\partial f}{\partial x}\right|_{(x_0,y_0)} (x - x_0) + \left.\frac{\partial f}{\partial y}\right|_{(x_0,y_0)} (y - y_0)\,, \tag{4.30}$$

which, obviously and by necessity, agrees with (4.27).

This discussion of the $n = 2$ case, notably (4.27) and (4.28), motivates the following definition of differentiability and the tangent map.

Definition: For $n \in \mathbb{N}$, let $f : A \to \mathbb{R}$ for $A \subset \mathbb{R}^n$ and let $\mathbf{x}_0 = (x_{01}, x_{02}, \ldots, x_{0n})'$ be an interior point of A. The function f is said to be *differentiable at* $\mathbf{x}_0$ if

1. $(\operatorname{grad} f)(\mathbf{x_0})$ exists, and
2. there exists a *tangent map* $T : \mathbb{R}^n \to \mathbb{R}$ of f at $\mathbf{x}_0$, such that

$$f(\mathbf{x}_0) = T(\mathbf{x}_0) \quad \text{and} \quad \lim_{\mathbf{x}\to\mathbf{x}_0} \frac{f(\mathbf{x}) - T(\mathbf{x})}{\|\mathbf{x} - \mathbf{x}_0\|} = 0, \tag{4.31}$$

where

$$T(\mathbf{x}) = f(\mathbf{x}_0) + (\operatorname{grad} f)(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0). \tag{4.32}$$

Formally, the definition does not include the above definition of T in terms of the gradient. However, if the tangent map of f at $\mathbf{x}_0$ exists, then it is unique, and given by (4.32). With $\mathbf{x}$ and $\mathbf{x}_0$ column vectors, (and recall grad is a row vector), we can write (4.32) as

$$T(\mathbf{x}) = f(\mathbf{x}_0) + \sum_{i=1}^{n} (D_i f)(\mathbf{x}_0)(x_i - x_{0i}) \tag{4.33}$$

$$=: f(\mathbf{x}_0) + \mathrm{d}f(\mathbf{x}_0, \mathbf{x} - \mathbf{x}_0), \tag{4.34}$$

where the term $\mathrm{d}f(\mathbf{x}_0, \mathbf{x} - \mathbf{x}_0)$ defined in (4.34) is the *total differential of* f *at* $\mathbf{x}_0$, i.e.,

$$\mathrm{d}f(\mathbf{x}, \mathbf{h}) = (\operatorname{grad} f)(\mathbf{x}) \cdot \mathbf{h}. \tag{4.35}$$

If f is differentiable at all points of A, then f is said to be *differentiable (on A)*.

We now need to explain why the limit condition for the tangent map in (4.28) for the $n = 2$ case, and, more generally, in (4.31), divides by $\|\mathbf{x} - \mathbf{x}_0\|$. To do so, we review the $n = 1$ case: Recall from (2.287) and (2.288) with $x_0 = c$ that, for $f : D \subset \mathbb{R} \to \mathbb{R}$, $x_0 \in D^o$ (the interior of D), and f such that, $\forall x \in D^o$, $\exists f''(x)$,

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + r(x), \quad r(x) = \frac{1}{2} f''(\zeta)\,(x - x_0)^2\,, \tag{4.36}$$

with

$$\lim_{x\to x_0} \frac{r(x)}{|x - x_0|} = \frac{1}{2} f''(\zeta) \lim_{x\to x_0} \frac{(x - x_0)^2}{|x - x_0|} = \frac{1}{2} f''(\zeta) \lim_{x\to x_0} |x - x_0| = 0. \tag{4.37}$$

This means $f(x_0) + f'(x_0)(x - x_0)$ is an affine linear approximation to $f(x)$ with the property that, not only is the error term $r(x)$ such that $\lim_{x\to x_0} r(x) = 0$, but also the limit of $r(x)$ after dividing by the linear quantity that itself goes to zero, is zero.

Remark: Another way of explaining the stronger condition of dividing by $\|\mathbf{x}-\mathbf{x}_0\|$ in (4.31) is given at the end of this subsection, in the excerpt from Lang's book. He also starts with the $n=1$ case, as we did above. We include this "redundancy" because of the importance of this tangent map criterion; and because it often helps to see the same idea presented in slightly different ways. Furthermore, the tangent map is a first-order approximation to the function, and below, in §4.8, we will define and use kth order approximations, these resulting from the Taylor series expansion of the function. The reader can quickly peak ahead and look at equation (4.122), and also the expression of (4.31) in terms of what is called the First-Order Approximation Theorem, in (4.123). Indeed, the first half of §4.8 is yet another description and development of (4.31), taken from Fitzpatrick's book. Its equation (4.144) gives the generalization of error term $r(x)$ in (4.36) to the general n case. The second half of that subsection details the second-order approximation, and its relevance for determining minima and maxima of functions, generalizing the univariate results (2.74) and (2.75).

The obvious generalization of (4.36) and (4.37) to the $n=2$ case of the differentiability condition (4.31) is, for $f: D \subset \mathbb{R}^2 \to \mathbb{R}$ and $(x_0, y_0) \in D^o$, there exists a real function $r(x)$ such that, $\forall (x,y) \in D^o$,

$$f(x,y) = f(x_0,y_0) + (D_1 f(x_0,y_0))(x-x_0) + (D_2 f(x_0,y_0))(y-y_0) + r(x,y), \tag{4.38}$$

with

$$\lim_{(x,y)\to(x_0,y_0)} \frac{r(x,y)}{\sqrt{(x-x_0)^2+(y-y_0)^2}} = 0. \tag{4.39}$$

For general n, with $f: D \subset \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{x}, \mathbf{x}_0 \in D^o$, and using definition (4.32),

$$f(\mathbf{x}) = T(\mathbf{x}) + r(\mathbf{x}) = f(\mathbf{x}_0) + (\text{grad}\, f)(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) + r(\mathbf{x}), \tag{4.40}$$

with, from (4.31),

$$r(\mathbf{x}) = f(\mathbf{x}) - T(\mathbf{x}), \qquad \lim_{\mathbf{x}\to\mathbf{x}_0} \frac{r(\mathbf{x})}{\|\mathbf{x}-\mathbf{x}_0\|} = 0. \tag{4.41}$$

The latter term in (4.41) obviously implies

$$\lim_{\mathbf{x}\to\mathbf{x}_0} r(\mathbf{x}) = 0. \tag{4.42}$$

Theorem: If a function $f: \mathcal{O} \subset \mathbb{R}^n \to \mathbb{R}$ for open domain $\mathcal{O}$ is differentiable at $\mathbf{x}_0 \in \mathcal{O}$, then it is continuous at $\mathbf{x}_0$.

Proof: Recall the equivalent definitions of continuity in (4.10) and (4.11). The result is very clear for $n=2$ using $\mathbf{x}_0 = (x_0, y_0)$ in (4.38) and (4.39), so that, with $\mathbf{x} = (x,y)$,

$$\lim_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x}) = f(\mathbf{x}_0). \tag{4.43}$$

The general n case follows directly from (4.40) and (4.41).

Observe how the above definition of differentiability at a point $\mathbf{x}_0$ in the (interior of the) domain requires not just the existence of $(\text{grad}\, f)(\mathbf{x}_0)$, but also the tangent map (4.31). This dual condition implies that we could have existence of $(\text{grad}\, f)(\mathbf{x}_0)$, but not (4.31). This is true, as the following example shows.

Example 4.12 *(Petrovic, p. 344) Prove that the function*

$$f(x,y) = \begin{cases} x+y, & \text{if } x=0 \text{ or } y=0 \\ 1, & \text{otherwise} \end{cases}$$

has partial derivatives at $(0,0)$ *but there is no tangent map at this point.*

Solution. By definition,

$$\frac{\partial f}{\partial x}(0,0) = \lim_{h\to 0} \frac{f(h,0)-f(0,0)}{h} = \lim_{h\to 0} \frac{h}{h} = 1,$$

and, similarly, $(\partial f/\partial y)(0,0) = 1$*. However,* f *is not continuous at* $(0,0)$*. Indeed,* $f(0,0) = 0$ *, but for any* $n \in \mathbb{N}$*,* $f(1/n, 1/n) = 1$ *so* $\lim f(1/n, 1/n) = 1$*. This is bad news because, if we substitute* $(x_0, y_0) = (0,0)$ *and* $(x,y) = (1/n, 1/n)$ *in* (4.38) *and assume* (4.39)*, we would get*

$$f\left(\frac{1}{n}, \frac{1}{n}\right) = 0 + 1\left(\frac{1}{n} - 0\right) + 1\left(\frac{1}{n} - 0\right) + r\left(\frac{1}{n}, \frac{1}{n}\right)$$

The left side equals 1, but the right side converges to 0. Thus, (4.38) *and* (4.39) *do not hold, there is no tangent map at this point, and* f *is not differentiable.* ■

In light of the previous example, in which f was not continuous, one might ask: If $(\operatorname{grad} f)(\mathbf{x}_0)$ exists, and f is continuous, then does the tangent map (4.31) exist? The answer is no, as the next example shows.

Example 4.13 *(Petrovic, p. 345) Prove that the function*

$$f(x,y) = \begin{cases} \frac{x^2 y}{x^2+y^2}, & \text{if } x^2+y^2 > 0, \\ 0, & \text{if } (x,y) = (0,0), \end{cases}$$

has partial derivatives at $(0,0)$*, and it is continuous at* $(0,0)$*, but* (4.28) *does not hold.*

Solution. This time f *is continuous. It is easy to see that this is true at any point different from the origin. For the continuity at the origin, we will show that*

$$\lim_{(x,y)\to(0,0)} \frac{x^2 y}{x^2+y^2} = 0. \tag{4.44}$$

Using the arithmetic-geometric mean inequality $|2xy| \le x^2 + y^2$*, we obtain that*

$$0 \le \left|\frac{x^2 y}{x^2+y^2}\right| \le \left|\frac{x^2 y}{2xy}\right| = \frac{|x|}{2},$$

which implies (4.44) *via the Squeeze Theorem* (2.3)*. So,* f *is continuous. Also, the partial derivatives at* $(0,0)$ *exist:*

$$\frac{\partial f}{\partial x}(0,0) = \lim_{h\to 0} \frac{f(h,0)-f(0,0)}{h} = \lim_{h\to 0} \frac{0}{h} = 0,$$

and, similarly, $(\partial f/\partial y)(0,0) = 0$*. However,* (4.28) *does not hold. Otherwise, we would have* $f(x,y) = r(x,y)$*, and it would follow that*

$$\lim_{(x,y)\to(0,0)} \frac{f(x,y)}{\sqrt{(x)^2+(y)^2}} = 0.$$

In particular, taking once again $(x,y)=(1/n,1/n)$*, we would obtain that*

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

but this is incorrect because the limit on the left side is $1/(2\sqrt{2})$*.* ■

Sufficient conditions for $f:D\subset\mathbb{R}^n\to\mathbb{R}$, with D open, in order for the tangent map (4.34) to exist at $\mathbf{x}_0\in D$, are:

$$\text{(i) the existence of grad } f \text{ on } D; \qquad \text{and} \qquad \text{(ii) continuity of grad } f \text{ at } \mathbf{x}_0.$$

These are fulfilled if f is continuously differentiable, i.e., grad f (exists and) is continuous (on D), denoted $f\in\mathcal{C}^1(D)$. The proof, from Petrovic, p. 345, is for $n=2$, with the general $n\in\mathbb{N}$ case clear in principle.

Theorem: Let $f:D\subset\mathbb{R}^2\to\mathbb{R}$ with D open. Suppose that partial derivatives $\partial f/\partial x$ and $\partial f/\partial y$ exist in D and that they are continuous at $(x_0,y_0)\in D$. Then

$$f \textbf{ is differentiable at } (x_0,y_0). \tag{4.45}$$

Proof: We will start with the equality

$$f(x,y)-f(x_0,y_0)=[f(x,y)-f(x_0,y)]+[f(x_0,y)-f(x_0,y_0)].$$

The existence of partial derivatives allows us to apply the Mean Value Theorem to each pair above. We obtain that

$$f(x,y)-f(x_0,y_0)=\frac{\partial f}{\partial x}(z,y)(x-x_0)+\frac{\partial f}{\partial y}(x_0,w)(y-y_0),$$

for some real numbers z (between x and x_0) and w (between y and y_0). We will write

$$\frac{\partial f}{\partial x}(z,y)=\frac{\partial f}{\partial x}(x_0,y_0)+\alpha,\quad \frac{\partial f}{\partial y}(x_0,w)=\frac{\partial f}{\partial y}(x_0,y_0)+\beta,$$

and the continuity of partial derivatives at (x_0,y_0) implies that, when $(x,y)\to(x_0,y_0)$, $\alpha,\beta\to 0$. Therefore,

$$\begin{aligned}f(x,y)-f(x_0,y_0)&=\left(\frac{\partial f}{\partial x}(x_0,y_0)+\alpha\right)(x-x_0)+\left(\frac{\partial f}{\partial y}(x_0,y_0)+\beta\right)(y-y_0)\\&=\frac{\partial f}{\partial x}(x_0,y_0)(x-x_0)+\frac{\partial f}{\partial y}(x_0,y_0)(y-y_0)+\alpha(x-x_0)+\beta(y-y_0).\end{aligned}$$

This is precisely the form (4.38), and the result will follow if we can show (4.39), i.e.,

$$\lim_{(x,y)\to(x_0,y_0)}\frac{\alpha(x-x_0)+\beta(y-y_0)}{\sqrt{(x-x_0)^2+(y-y_0)^2}}=0.$$

Notice that

$$\frac{|x-x_0|}{\sqrt{(x-x_0)^2+(y-y_0)^2}},\frac{|y-y_0|}{\sqrt{(x-x_0)^2+(y-y_0)^2}}\le 1.$$

It follows that

$$0\le\left|\frac{\alpha(x-x_0)+\beta(y-y_0)}{\sqrt{(x-x_0)^2+(y-y_0)^2}}\right|\le|\alpha|+|\beta|\to 0,\quad (x,y)\to(x_0,y_0).$$

Note that, as stated above, the theorem provides a set of sufficient conditions. It is thus not true that differentiability of f, i.e., the existence of grad f and the existence of the tangent map (4.34), implies that any or all of the $(D_i f)$ are continuous.

The proof for the $n = 3$ case starts the same, and requires writing $f(x, y, z) - f(x_0, y_0, z_0)$ as the appropriate sum of three terms. This sum is given below in (4.71) and (4.72). The rest of the proof is then the same. Although cumbersome, one could attempt to write the relevant sum expansion of $f(\mathbf{x}) - f(\mathbf{x}_0)$ for the general n case.

Observe that, combining the theorems (4.43) and (4.45) (and assuming the latter holds for general n), we have that, for $f : \mathcal{O} \subset \mathbb{R}^n \to \mathbb{R}$,

$$f \in \mathcal{C}^1(\mathcal{O}) \implies f \in \mathcal{C}^0(\mathcal{O}), \quad \textbf{i.e., } f \textbf{ is continuous on } \mathcal{O}. \tag{4.46}$$

In §4.5, page 229, we give the proof of this result using the multivariate MVT.

We now turn to differentiability of basic functions of two differentiable functions, namely additivity and homogeneity; the dot product (4.15); and, as the range of f and g are $\mathbb{R}$, the quotient.

Let $f, g : D \subset \mathbb{R}^n \to \mathbb{R}$ be differentiable, and let $\mathbf{x}, \mathbf{x}_0 \in D^o$. From (4.40) and (4.31), but adding appropriate subscripts to T and r,

$$f(\mathbf{x}) = T_f(\mathbf{x}) + r_f(\mathbf{x}) = f(\mathbf{x}_0) + (\operatorname{grad} f)(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + r_f(\mathbf{x}), \tag{4.47}$$

$$g(\mathbf{x}) = T_g(\mathbf{x}) + r_g(\mathbf{x}) = g(\mathbf{x}_0) + (\operatorname{grad} g)(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + r_g(\mathbf{x}), \tag{4.48}$$

with $r_f(\mathbf{x}) = f(\mathbf{x}) - T_f(\mathbf{x})$, $r_g(\mathbf{x}) = g(\mathbf{x}) - T_g(\mathbf{x})$, and

$$\lim_{\mathbf{x}\to\mathbf{x}_0} \frac{r_f(\mathbf{x})}{\|\mathbf{x} - \mathbf{x}_0\|} = 0, \qquad \lim_{\mathbf{x}\to\mathbf{x}_0} \frac{r_g(\mathbf{x})}{\|\mathbf{x} - \mathbf{x}_0\|} = 0. \tag{4.49}$$

Theorem: Under the above conditions on f, g, and with $k, k_1, k_2 \in \mathbb{R}$,

$$k_1 f + k_2 g \textbf{ is differentiable,} \tag{4.50}$$

with

$$\operatorname{grad}(k_1 f + k_2 g) = k_1 \operatorname{grad}(f) + k_2 \operatorname{grad}(g). \tag{4.51}$$

Notice differentiable functions form a vector space.

Proof: For homogeneity, analogous to the linearity property of differentiation in the univariate case, we see that, from (4.20), with $(kf)(\mathbf{x}) = kf(\mathbf{x})$, $(D_i(kf))(\mathbf{x}) = k(D_i f)(\mathbf{x})$. Thus, from (4.22), $(\nabla(kf))(\mathbf{x}) = (\operatorname{grad}(kf))(\mathbf{x}) = k\,(D_1 f(\mathbf{x}), \ldots, D_n f(\mathbf{x}))$, and, simply multiplying (4.47) by k,

$$(kf)(\mathbf{x}) = kf(\mathbf{x}) = kf(\mathbf{x}_0) + k(\operatorname{grad} f)(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + kr_f(\mathbf{x}).$$

With $N_{\mathbf{x}} = \|\mathbf{x} - \mathbf{x}_0\|$, the linearity property (4.12) implies $\lim_{\mathbf{x}\to\mathbf{x}_0} kr_f(\mathbf{x})/N_{\mathbf{x}} = k \lim_{\mathbf{x}\to\mathbf{x}_0} r_f(\mathbf{x})/N_{\mathbf{x}} = 0$, so that (kf) is differentiable: $\operatorname{grad}(kf)$ exists and is $k \operatorname{grad} f$; and the tangent map exists and is $T_{(kf)} = kT_f$.

For linearity, adding (4.47) and (4.48), and using the univariate result (2.37), we have $(\operatorname{grad}(f + g))(\mathbf{x}_0) = (\operatorname{grad} f)(\mathbf{x}_0) + (\operatorname{grad} g)(\mathbf{x}_0)$ and $r_{f+g}(\mathbf{x}) = r_f(\mathbf{x}) + r_g(\mathbf{x})$. It remains to show that $\lim_{\mathbf{x}\to\mathbf{x}_0} r_{f+g}(\mathbf{x})/N_{\mathbf{x}} = 0$. But this again follows from (4.49) and the linearity property of limits (4.12).

Theorem: Let $f, g : A \subset \mathbb{R}^n \to \mathbb{R}$ be differentiable at $\mathbf{a} \in A^o$. Then $f \cdot g$ is differentiable at $\mathbf{a}$, and

$$\operatorname{grad}(f \cdot g)(\mathbf{a}) = \operatorname{grad} f(\mathbf{a})g(\mathbf{a}) + f(\mathbf{a}) \operatorname{grad} g(\mathbf{a}). \tag{4.52}$$

If, in addition, $g(\mathbf{a}) \neq 0$, then the function f/g is differentiable at $\mathbf{a}$, and

$$\operatorname{grad}\left(\frac{f}{g}\right)(\mathbf{a}) = \frac{\operatorname{grad} f(\mathbf{a})g(\mathbf{a}) - f(\mathbf{a})\mathbf{D}g(\mathbf{a})}{g(\mathbf{a})^2}. \tag{4.53}$$

Proof: For convenience, denote $(\operatorname{grad} f)(\mathbf{x}) = (D_1 f(\mathbf{x}), \ldots, D_n f(\mathbf{x}))$ as $(A_1, \ldots, A_n)$. Likewise, let $(\operatorname{grad} g)(\mathbf{x}) = (B_1, \ldots, B_n)$. For (4.52), multiplying (4.47) and (4.48), and with $\mathbf{x} \in A^o$,

$$\begin{aligned} f(\mathbf{x})g(\mathbf{x}) = f(\mathbf{a})g(\mathbf{a}) &+ \sum_{i=1}^{n} \left(A_i g(\mathbf{a}) + B_i f(\mathbf{a})\right)(x_i - a_i) + \sum_{i=1}^{n}\sum_{j=1}^{n} A_i B_j (x_i - a_i)(x_j - a_j) \\ &+ r_f(\mathbf{x})g(\mathbf{x}) + r_g(\mathbf{x})f(\mathbf{x}) - r_f(\mathbf{x})r_g(\mathbf{x}). \end{aligned}$$

Consider the last three remainder terms. From the basic multiplicity property of limits and the differentiability of f; in particular (4.49),

$$\lim_{\mathbf{x}\to\mathbf{a}} \frac{r_f(\mathbf{x})g(\mathbf{x})}{\|\mathbf{x}-\mathbf{a}\|} = \lim_{\mathbf{x}\to\mathbf{a}} g(\mathbf{x}) \cdot \lim_{\mathbf{x}\to\mathbf{a}} \frac{r_f(\mathbf{x})}{\|\mathbf{x}-\mathbf{a}\|} = g(\mathbf{a}) \cdot 0 = 0,$$

and likewise for the $r_g(\mathbf{x})f(\mathbf{x})$ term. For the $r_f(\mathbf{x})r_g(\mathbf{x})$ term, from (4.42),

$$\lim_{\mathbf{x}\to\mathbf{a}} \frac{r_f(\mathbf{x})r_g(\mathbf{x})}{\|\mathbf{x}-\mathbf{a}\|} = \lim_{\mathbf{x}\to\mathbf{a}} r_f(\mathbf{x}) \cdot \lim_{\mathbf{x}\to\mathbf{a}} \frac{r_g(\mathbf{x})}{\|\mathbf{x}-\mathbf{a}\|} = 0 \cdot 0.$$

For the double sum in the above product expression, note that each of its n^2 terms satisfies

$$0 \leq \frac{|x_i - a_i|\,|x_j - a_j|}{\|\mathbf{x}-\mathbf{a}\|} \leq |x_i - a_i|,$$

so that taking limits and using the Squeeze Theorem (2.3) implies that each term, and thus the double sum, divided by $\|\mathbf{x}-\mathbf{a}\|$, converges to zero.

Thus, we can define $r_{f\cdot g}$ to be the sum of the last four components in the above product expression for $f(\mathbf{x})g(\mathbf{x})$. Using (4.47) and (4.48) as analogies, the single-term sum in the above expression for $f(\mathbf{x})g(\mathbf{x})$ must be $(\operatorname{grad}(f \cdot g)(\mathbf{a})(\mathbf{x} - \mathbf{a})$. This can be expressed as

$$\begin{aligned} \operatorname{grad}(f \cdot g)(\mathbf{a}) &= \begin{bmatrix} A_1 g(\mathbf{a}) + B_1 f(\mathbf{a}) & A_2 g(\mathbf{a}) + B_2 f(\mathbf{a}) & \ldots & A_n g(\mathbf{a}) + B_n f(\mathbf{a}) \end{bmatrix} \\ &= g(\mathbf{a}) \begin{bmatrix} A_1 & A_2 & \ldots & A_n \end{bmatrix} + f(\mathbf{a}) \begin{bmatrix} B_1 & B_2 & \ldots & B_n \end{bmatrix} \\ &= g(\mathbf{a})\mathbf{D}f(\mathbf{a}) + f(\mathbf{a})\mathbf{D}g(\mathbf{a}). \end{aligned}$$

For the quotient result (4.53), see Petrovic, p. 349.

I now give a different (and far less efficient; and possibly faulty) proof than the one above, around (4.43), for the $n = 2$ case, that differentiability of $f : A \subset \mathbb{R}^2 \to \mathbb{R}$, A open, implies f is continuous (on A). The goal was to use the ϵ-δ formulation of continuity in (4.18) instead of the (in this case, far easier) sequential limit formulation (4.17). The right way to do the proof is shown in (4.92), which is for general n and m.

The odd thing about my attempt is that it appears to shown $\lim_{(\delta_x,\delta_y)\to(0,0)} R_{x,y} = 0$, which is the requirement in (4.39). In other words, it appears as though only the existence of grad f was required, as opposed to also requiring existence of the tangent map.

Recall the fundamental lemma of differentiation (2.35):

> Let $f : A \subset \mathbb{R} \to \mathbb{R}$ be differentiable at the point x. Then there exists a function η defined on an interval about zero such that
>
> $$f(x+h) - f(x) = \left[f'(x) + \eta(h)\right] \cdot h, \tag{4.54}$$
>
> and η is continuous at zero, with $\eta(0) = 0$.

"Proof": Let $f : A \subset \mathbb{R}^2 \to \mathbb{R}$ be differentiable on the open set A. We wish to show that f is also continuous on A. For $(x, y) \in A$, let $z = f(x, y)$ and let δ_x and δ_y represent very small, positive quantities such that $(x + \delta_x, y + \delta_y) \in A$. Let δ_z be such that $z + \delta_z = f(x + \delta_x, y + \delta_y)$, i.e.,

$$\begin{aligned}
\delta_z &= f(x+\delta_x, y+\delta_y) - f(x,y) \\
&= f(x+\delta_x, y+\delta_y) - f(x, y+\delta_y) + f(x, y+\delta_y) - f(x,y) \\
&= \frac{f(x+\delta_x, y+\delta_y) - f(x, y+\delta_y)}{\delta_x}\delta_x + \frac{f(x, y+\delta_y) - f(x,y)}{\delta_y}\delta_y.
\end{aligned} \tag{4.55}$$

Using (4.54), with δ_x playing the role of h in (4.54); and, separately, δ_y playing the role of h in (4.54), (4.55) can be written as

$$\begin{aligned}
\delta_z &= \left(\frac{\partial f(x, y+\delta_y)}{\partial x} + \eta_x\right)\delta_x + \left(\frac{\partial f(x,y)}{\partial y} + \eta_y\right)\delta_y \\
&= \frac{\partial f(x, y+\delta_y)}{\partial x}\delta_x + \frac{\partial f(x,y)}{\partial y}\delta_y \;+\; \eta_x\delta_x \;+\; \eta_y\delta_y,
\end{aligned} \tag{4.56}$$

where the existence of $D_1 f$ and $D_2 f$ was assumed via differentiability, and which is of the form (4.38) with $r(\mathbf{x}) = \eta_x\delta_x + \eta_y\delta_y$. With

$$N_{x,y} := \sqrt{\delta_x^2 + \delta_y^2}, \qquad \text{and} \qquad R_{x,y} := \left|\frac{\eta_x\delta_x + \eta_y\delta_y}{N_{x,y}}\right|,$$

and using the triangle inequality,

$$0 \le \lim_{(\delta_x,\delta_y)\to(0,0)} R_{x,y} \le \lim_{(\delta_x,\delta_y)\to(0,0)} |\eta_x|\frac{|\delta_x|}{N_{x,y}} + |\eta_y|\frac{|\delta_y|}{N_{x,y}} < \lim_{(\delta_x,\delta_y)\to(0,0)} |\eta_x| + |\eta_y|, \tag{4.57}$$

because $0 < |\delta_x|/N_{x,y} < 1$ and $0 < |\delta_y|/N_{x,y} < 1$. The continuity of the η function and that $\eta(0) = 0$ implies $\lim_{\delta_x\to 0}\eta_x = \lim_{\delta_y\to 0}\eta_y = 0$. Thus, the Squeeze Theorem (2.3) implies the lhs limit in (4.57) is zero; which obviously implies

$$\lim_{(\delta_x,\delta_y)\to(0,0)} (\eta_x\delta_x + \eta_y\delta_y) = 0.$$

That is, $\exists\delta_x > 0$ and $\exists\delta_y > 0$ such that δ_z in (4.56) can be made arbitrarily close to zero. Thus, as $(x, y) \in A$ was arbitrary, f is continuous on A.

For some enrichment, we give the presentation from Lang (1987, §III.3), showing the extension of (4.54) to the $n > 1$ case.

Let f be a function defined on an open set U. Let X be a point of U. For all vectors H such that $\|H\|$ is small (and $H \neq O$), the point $X + H$ also lies in the open set. However, we cannot form a quotient

$$\frac{f(X+H) - f(x)}{H},$$

because it is meaningless to divide by a vector. In order to define what we mean for a function f to be differentiable, we must therefore find a way that does not involve dividing by H. We reconsider the case of functions of one variable. Let us fix a number x. We had defined the derivative to be

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}.$$

Let

$$\varphi(h) = \frac{f(x+h) - f(x)}{h} - f'(x).$$

Then $\varphi(h)$ is not defined when $h = 0$, but $\lim_{h \to 0} \varphi(h) = 0$. We can write

$$f(x+h) - f(x) = f'(x)h + h\varphi(h).$$

This relation has meaning so far only when $h \neq 0$. However, we observe that if we define $\varphi(0)$ to be 0, then the preceding relation is obviously true when $h = 0$ (because we just get $0 = 0$). Let

$$g(h) = \varphi(h), \quad \text{if} \quad h > 0, \qquad g(h) = -\varphi(h), \quad \text{if} \quad h < 0.$$

Then we have shown that if f is differentiable, there exists a function g such that

$$f(x+h) - f(x) = f'(x)h + |h|g(h), \tag{4.58}$$

and $\lim_{h \to 0} g(h) = 0$. Conversely, suppose that there exists a number a and a function $g(h)$ such that

$$f(x+h) - f(x) = ah + |h|g(h) \tag{4.59}$$

and

$$\lim_{h \to 0} g(h) = 0.$$

We find for $h \neq 0$,

$$\frac{f(x+h) - f(x)}{h} = a + \frac{|h|}{h} g(h).$$

Taking the limit as h approaches 0, we observe that

$$\lim_{h \to 0} \frac{|h|}{h} g(h) = 0.$$

Hence, the limit of the Newton quotient exists and is equal to a. Hence f is differentiable, and its derivative $f'(x)$ is equal to a.

Therefore, the existence of a number a and a function g satisfying (4.59) could have been used as the definition of differentiability in the case of functions of one variable. The great advantage of (4.58) is that no h appears in the denominator. It is this relation that

will suggest to us how to define differentiability for functions of several variables, and how to prove the chain rule for them.

Let us begin with two variables. We let $X = (x, y)$ and $H = (h, k)$. Then the notion corresponding to $x+h$ in one variable is here $X+H = (x+h, y+k)$. We wish to compare the values of a function f at X and $X + H$, i.e. we wish to investigate the difference

$$f(X + H) - f(X) = f(x + h, y + k) - f(x, y).$$

Definition: We say that f is *differentiable* at X if the partial derivatives

$$\frac{\partial f}{\partial x} \quad \text{and} \quad \frac{\partial f}{\partial y}$$

exist, and if there exists a function g (defined for small H) such that $\lim_{H\to O} g(H) = 0$ and

$$f(x + h, y + k) - f(x, y) = \frac{\partial f}{\partial x}h + \frac{\partial f}{\partial y}k + \|H\|g(H). \tag{4.60}$$

We view the term

$$\frac{\partial f}{\partial x}h + \frac{\partial f}{\partial y}k$$

as an approximation to $f(X + H) - f(X)$, depending in a particularly simple way on h and k. If we use the abbreviation $\operatorname{grad} f = \nabla f$, then (4.60) can be written

$$f(X + H) - f(X) = \nabla f(x) \cdot H + \|H\|g(H).$$

As with $\operatorname{grad} f$, one must read $(\nabla f)(X)$ and not the meaningless $\nabla(f(X))$ since $f(X)$ is a number for each value of X, and thus it makes no sense to apply ∇ to a number. The symbol ∇ is applied to the function f, and $(\nabla f)(X)$ is the value of ∇f at X.

We now consider a function of n variables. Let f be a function defined on an open set U. Let X be a point of U. If $H = (h_1, \ldots, h_n)$ is a vector such that $\|H\|$ is small enough, then $X + H$ will also be a point of U and so $f(X + H)$ is defined. Note that

$$X + H = (x_1 + h_1, \ldots, x_n + h_n)\,.$$

This is the generalization of the $x + h$ with which we dealt previously in one variable, or the $(x + h, y + k)$ in two variables. For three variables, we already run out of convenient letters, so we may as well write n instead of 3.

Definition: We say that f is differentiable at X if, first, all the partial derivatives $D_i f(X)$ exist, $i = 1, \ldots, n$; and, second, if there exists a function g (defined for small H) such that $\lim_{H\to O} g(H) = 0$, also written $\lim_{\|H\|\to 0} g(H) = 0$, whereby

$$f(X + H) - f(X) = D_1 f(X)h_1 + \cdots + D_n f(x)h_n + \|H\|g(H).$$

We say that f is *differentiable in the open set* U if it is differentiable at every point of U, so that the above relation holds for every point $X \in U$. In view of the definition of the gradient, we can rewrite our fundamental relation in the form

$$f(X + H) - f(X) = (\operatorname{grad} f(X)) \cdot H + \|H\|g(H).$$

The term $\|H\|g(H)$ has an order of magnitude smaller than the previous term involving the dot product. This is one advantage of the present notation. We know how to handle

the formalism of dot products and are accustomed to it, and its geometric interpretation. This will help us later in interpreting the gradient geometrically.

As an example, suppose that we consider values for H pointing only in the direction of the standard unit vectors. In the case of two variables, consider for instance $H = (h, 0)$. Then for such H, the condition for differentiability reads:

$$f(X + H) = f(x + h, y) = f(x, y) + \frac{\partial f}{\partial x}h + |h|g(H).$$

In higher dimensional space, let $E_i = (0, \ldots, 0, 1, 0, \ldots, 0)$ be the ith unit vector. Let $H = hE_i$ for some number h, so that $H = (0, \ldots, 0, h, 0, \ldots, 0)$. Then for such H,

$$f(X + H) = f\left(X + hE_i\right) = f(X) + \frac{\partial f}{\partial x_i}h + |h|g(H),$$

and, therefore, if $h \neq 0$, we obtain

$$\frac{f(X + H) - f(X)}{h} = D_i f(X) + \frac{|h|}{h}g(H).$$

Because of the special choice of H, we can divide by the number h, but we are not dividing by the vector H.

4.4 Higher Order Partial Derivatives

We now turn to second-order partial derivatives, using for this subsection the presentation in Fitzpatrick. Further comments I have added for clarity are indicated in blue color.

Given an open subset $\mathcal{O}$ of $\mathbb{R}^n$ and an index i with $1 \leq i \leq n$, if the function $f : \mathcal{O} \to \mathbb{R}$ has a partial derivative with respect to its ith component at each point in $\mathcal{O}$, then the function $\partial f/\partial x_i : \mathcal{O} \to \mathbb{R}$ is defined and we can ask whether this new function itself has first-order partial derivatives. Fix an index j with $1 \leq j \leq n$. If the function $\partial f/\partial x_i : \mathcal{O} \to \mathbb{R}$ has a partial derivative with respect to its jth component at the point $\mathbf{x}$ in $\mathcal{O}$, we write

$$\frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}) \quad \text{to denote} \quad \frac{\partial}{\partial x_j}\left[\frac{\partial f}{\partial x_i}\right](\mathbf{x}).$$

When $n = 2$ or 3, and points are labeled without subscripts, we use a more suggestive notation for second partial derivatives; e.g., $\partial^2 f/\partial x \partial y$, etc..

Definition: Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ and consider a function $f : \mathcal{O} \to \mathbb{R}$:

i. The function $f : \mathcal{O} \to \mathbb{R}$ is said to have second-order partial derivatives provided that it has first-order partial derivatives and that, for each index i with $1 \leq i \leq n$, the function $\partial f/\partial x_i : \mathcal{O} \to \mathbb{R}$ also has first-order partial derivatives.

ii. The function $f : \mathcal{O} \to \mathbb{R}$ is said to have continuous second-order partial derivatives provided that it has second-order partial derivatives and that, for each pair of indices i and j with $1 \leq i \leq n$ and $1 \leq j \leq n$, the function $\partial^2 f/\partial x_i \partial x_j : \mathcal{O} \to \mathbb{R}$ is continuous.

So, if f is continuously differentiable, it means its first derivatives (exist and) are continuous. The Continuity Theorem mentioned above then implies f is continuous. Apply this to the second derivatives: If second derivatives (exist and) are continuous, then applying the previous result, first derivatives are continuous, i.e., the function f is continuously differentiable.

We now state and prove a fundamental result regarding exchange of the partial derivative operator. Some books refer to this as Clairaut's, or Schwarz's, Theorem. We give the proof as in Fitzpatrick, but the reader can also see, e.g., Lang (1997, p. 372) or Protter and Morrey (1991, p. 179).

Theorem (Fitzpatrick, 13.10): Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ and suppose that the function $f : \mathcal{O} \to \mathbb{R}$ has continuous second-order partial derivatives. For any two indices i and j with $1 \leq i \leq n$ and $1 \leq j \leq n$ and any point $\mathbf{x}$ in $\mathcal{O}$,

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}). \tag{4.61}$$

To prove this, we first require the following lemma:

Lemma (Fitzpatrick, 13.11): Let $\mathcal{U}$ be an open subset of the plane $\mathbb{R}^2$ that contains the point (x_0, y_0) and suppose that the function $f : \mathcal{U} \to \mathbb{R}$ has second-order partial derivatives (not necessarily continuous). Then there are points (x_1, y_1) and (x_2, y_2) in $\mathcal{U}$ at which

$$\frac{\partial^2 f}{\partial x \partial y}(x_1, y_1) = \frac{\partial^2 f}{\partial y \partial x}(x_2, y_2).$$

Proof: Since $\mathcal{U}$ is open, we can choose a positive number r such that if we define the intervals of real numbers I and J by $I = (x_0 - 2r, x_0 + 2r)$ and $J = (y_0 - 2r, y_0 + 2r)$, then the rectangle $I \times J$ is contained in $\mathcal{U}$. The idea of the proof is to express

$$f(x_0 + r, y_0 + r) - f(x_0 + r, y_0) - f(x_0, y_0 + r) + f(x_0, y_0)$$

as a difference in two different ways: First as the difference

$$[f(x_0 + r, y_0 + r) - f(x_0 + r, y_0)] - [f(x_0, y_0 + r) - f(x_0, y_0)], \tag{4.62}$$

(Below, this is the same as $\phi(x_0 + r) - \phi(x_0)$.), and then as the difference

$$[f(x_0 + r, y_0 + r) - f(x_0, y_0 + r)] - [f(x_0 + r, y_0) - f(x_0, y_0)]. \tag{4.63}$$

Then we use the Mean Value Theorem for functions of a single real variable to express (4.62) and (4.63) as second-order partial derivatives of the function $f : \mathcal{U} \to \mathbb{R}$.

First we analyze the difference (4.62). Define the auxiliary function $\varphi : I \to \mathbb{R}$ by

$$\varphi(x) = f(x, y_0 + r) - f(x, y_0) \qquad \text{for } x \text{ in } I.$$

Since $f : \mathcal{U} \to \mathbb{R}$ has a partial derivative with respect to its first component, the function $\varphi : I \to \mathbb{R}$ is differentiable. Thus, we can apply the Mean Value Theorem to the restriction of the function $\varphi : I \to \mathbb{R}$ to the closed interval $[x_0, x_0 + r]$ to select a point x_1 in the open interval $(x_0, x_0 + r)$ such that

$$\frac{\varphi(x_0 + r) - \varphi(x_0)}{r} = \varphi'(x_1);$$

that is, (Below, this is $\alpha(y_0 + r) - \alpha(y_0)$.)

$$\frac{\varphi(x_0 + r) - \varphi(x_0)}{r} = \frac{\partial f}{\partial x}(x_1, y_0 + r) - \frac{\partial f}{\partial x}(x_1, y_0). \tag{4.64}$$

With this point x_1 fixed, define another auxiliary function $\alpha : J \to \mathbb{R}$ by

$$\alpha(y) = \frac{\partial f}{\partial x}(x_1, y), \quad y \in J.$$

We can apply the Mean Value Theorem to the restriction of the function $\alpha : J \to \mathbb{R}$ to the closed interval $[y_0, y_0 + r]$ to select a point y_1 in the open interval $(y_0, y_0 + r)$ such that

$$\frac{\alpha(y_0 + r) - \alpha(y_0)}{r} = \frac{\partial^2 f}{\partial y \partial x}(x_1, y_1), \quad x_1 \in (x_0, x_0 + r), \;\; y_1 \in (y_0, y_0 + r). \tag{4.65}$$

From (4.64) and (4.65), we obtain

$$\varphi(x_0 + r) - \varphi(x_0) = r^2 \frac{\partial^2 f}{\partial y \partial x}(x_1, y_1).$$

However, $\varphi(x_0 + r) - \varphi(x_0)$ equals the difference (4.62), and hence we have

$$\begin{aligned}&[f(x_0 + r, y_0 + r) - f(x_0 + r, y_0)] - [f(x_0, y_0 + r) - f(x_0, y_0)] \\ &\quad = r^2 \frac{\partial^2 f}{\partial y \partial x}(x_1, y_1).\end{aligned} \tag{4.66}$$

In order to analyze the difference (4.63), we now repeat the same argument applied to the auxiliary function $\psi : J \to \mathbb{R}$ defined by

$$\psi(y) = f(x_0 + r, y) - f(x_0, y) \quad y \in J.$$

From this it will follow that we can select a point (x_2, y_2) in the rectangle $I \times J$ such that

$$\begin{aligned}&[f(x_0 + r, y_0 + r) - f(x_0, y_0 + r)] - [f(x_0 + r, y_0) - f(x_0, y_0)] \\ &\quad = r^2 \frac{\partial^2 f}{\partial x \partial y}(x_2, y_2). \end{aligned} \tag{4.67}$$

From the equality of the left-hand sides of (4.66) and (4.67) follows the equality of the right-hand sides, so the lemma is proved.

Proof of Fitzpatrick Theorem 13.10: We prove the theorem when $n = 2$ and leave the general case to the reader. Let (x_0, y_0) be a point in $\mathcal{O}$. Choose a positive number r such that the open ball $\mathcal{B}_r(x_0, y_0)$ is contained in $\mathcal{O}$. Let k be a natural number. Then we can apply the lemma with $\mathcal{U} = \mathcal{B}_{r/k}(x_0, y_0)$ and select points (x_k, y_k) and (u_k, v_k) in $\mathcal{B}_{r/k}(x_0, y_0)$ at which

$$\frac{\partial^2 f}{\partial x \partial y}(x_k, y_k) = \frac{\partial^2 f}{\partial y \partial x}(u_k, v_k). \tag{4.68}$$

But, by assumption, the function $\partial^2 f / \partial x \partial y : \mathcal{O} \to \mathbb{R}$ is continuous at (x_0, y_0), as is the function $\partial^2 f / \partial y \partial x : \mathcal{O} \to \mathbb{R}$. Since the sequences $\{(x_k, y_k)\}$ and $\{(u_k, v_k)\}$ both converge to the point (x_0, y_0), it follows from (4.16) that

$$\lim_{k\to\infty}\left[\frac{\partial^2 f}{\partial x \partial y}(x_k, y_k)\right] = \frac{\partial^2 f}{\partial x \partial y}(x_0, y_0), \quad \lim_{k\to\infty}\left[\frac{\partial^2 f}{\partial y \partial x}(u_k, v_k)\right] = \frac{\partial^2 f}{\partial y \partial x}(x_0, y_0).$$

In view of these two equations, and taking limits in (4.68), we arrive at (4.61).

Observe that, in Lemma 13.11, we required only that the function $f : \mathcal{O} \to \mathbb{R}$ have second-order partial derivatives. On the other hand, in Theorem 13.10, we required that the second-order partial derivatives be continuous. This extra assumption is necessary. The following is an example of a function $f : \mathcal{O} \to \mathbb{R}$ that has second-order partial derivatives, and yet we do not have equality of $\partial^2 f / \partial x \partial y$ and $\partial^2 f / \partial y \partial x$ at all points.

Example 4.14 *(Fitzpatrick, p. 361) Define the function $f : \mathbb{R}^2 \to \mathbb{R}$ by*

$$f(x, y) = \begin{cases} xy(x^2 - y^2)/(x^2 + y^2), & \text{if } (x, y) \neq (0, 0), \\ 0, & \text{if } (x, y) = (0, 0). \end{cases}$$

Calculations show that the function $f : \mathbb{R}^2 \to \mathbb{R}$ has second-order partial derivatives but that

$$\frac{\partial^2 f}{\partial y \partial x}(0, 0) = -1 \quad \text{while} \quad \frac{\partial^2 f}{\partial x \partial y}(0, 0) = 1,$$

and thus not equal. ■

The next example shows an application of Clairaut's theorem (4.61).

Example 4.15 *(Bóna and Shabanov, Concepts in Calculus III, p. 153) Find $f(x,y,z)$ if $f'_x = yz + 2x = F_1$, $f'_y = xz + 3y^2 = F_2$, and $f'_z = xy + 4z^3 = F_3$; or show that it does not exist.*

Solution: The integrability conditions $(F_1)'_y = (F_2)'_x$, $(F_1)'_z = (F_3)'_x$, and $(F_2)'_z = (F_3)'_z$ are satisfied (their verification is left to the reader). So f exists. Taking the antiderivative with respect to x in the first equation, one finds

$$f'_x = yz + 2x \quad \Longrightarrow \quad f(x,y,z) = \int (yz + 2x)dx = xyz + x^2 + g(y,z),$$

where $g(y,z)$ is arbitrary. The substitution of f into the second equations yields

$$\begin{aligned} f'_y = xz + 3y^2 &\Longrightarrow xz + g'_y(y,z) = xz + 3y^2 \\ &\Longrightarrow g'_y(y,z) = 3y^2 \\ &\Longrightarrow g(y,z) = \int 3y^2 dy = y^3 + h(z) \\ &\Longrightarrow f(x,y,z) = xyz + x^2 + y^3 + h(z), \end{aligned}$$

where $h(z)$ is arbitrary. The substitution of f into the third equation yields

$$\begin{aligned} f'_z = xy + 4z^3 &\Longrightarrow xy + h'(z) = xy + 4z^3 \\ &\Longrightarrow h'(z) = 4z^3 \\ &\Longrightarrow h(z) = z^4 + c \\ &\Longrightarrow f(x,y,z) = xyz + x^2 + y^3 + z^4 + c, \end{aligned}$$

where c is a constant. ■

4.5 Directional Derivatives and the Multivariate MVT

This subsection, taken nearly verbatim from Fitzpatrick (2009, §13.3), is core material. For example, directional derivatives are fundamental in understanding multivariate function optimization. Whether minimizing a cost function or a financial risk measure; or maximizing an economic utility function or the statistical likelihood associated with a data set, optimization is perhaps the best example of the power and necessity of understanding this material. That holds, perhaps obviously, for gradient- and Hessian-based optimization algorithms, notably the very popular BFGS algorithm, and stochastic gradient descent in large-scale models in machine learning, but also for methods that are applicable for functions that are not differentiable, and perhaps not even continuous, such as evolutionary algorithms (differential evolution, genetic programming), tabu search, particle swarm, simulated annealing and, arguably the best of them all, CMA-ES, the latter also deeply intertwined with probability and statistical theory.[29]

A further goal of working through the material is to get to the proof that, if $f : \mathcal{O} \subset \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable on its open domain, then it is continuous. We have seen this result already in (4.46), though the proof of one of its parts, (4.45), was only heuristically determined for $n > 2$.

Further comments I have added for clarity are indicated in blue color.

Lemma (Fitzpatrick, 13.14, The Mean Value Lemma): Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ and let i be an index with $1 \leq i \leq n$. Suppose that the function $f : \mathcal{O} \to \mathbb{R}$ has a partial derivative with respect to its ith component at each point in $\mathcal{O}$. Let $\mathbf{x}$ be a point in $\mathcal{O}$ and let a be a real number such that the segment between the points $\mathbf{x}$ and $\mathbf{x} + a\mathbf{e}_i$ lies in $\mathcal{O}$. Then there is a number θ with $0 < \theta < 1$ such that

$$f(\mathbf{x} + a\mathbf{e}_i) - f(\mathbf{x}) = \frac{\partial f}{\partial x_i}(\mathbf{x} + \theta a \mathbf{e}_i)\, a. \tag{4.69}$$

Proof: Since $\mathcal{O}$ is open in $\mathbb{R}^n$, we can select an open interval of real numbers I that contains the numbers 0 and a such that, for each t in I, the point $\mathbf{x} + t\mathbf{e}_i$ belongs to $\mathcal{O}$.

It is very useful to first review the parametrized path formulation in the beginning of §4.2, notably equations (4.24) and (4.25).

Define the function $\phi : I \to \mathbb{R}$ by $\phi(t; i, \mathbf{x}) = \phi(t) = f(\mathbf{x} + t\mathbf{e}_i)$ for each t in I. Then the partial differentiability of the function $f : \mathcal{O} \to \mathbb{R}$ with respect to its ith component implies that, at each point t in I,

$$\phi'(t) = \frac{\partial f}{\partial x_i}(\mathbf{x} + t\mathbf{e}_i).$$

It follows that the function $\phi : I \to \mathbb{R}$ is differentiable. Thus, we can apply the Mean Value Theorem for functions of a single variable to the restriction of the function $\phi : I \to \mathbb{R}$ to the closed interval $[0, a]$ to obtain a point θ with $0 < \theta < 1$ such that

$$\phi(a) - \phi(0) = \phi'(\theta a)a,$$

which, in view of the definition of the function $\phi : I \to \mathbb{R}$ and the calculation of $\phi'(t)$, can be rewritten as (4.69).

[29] For an introduction to CMA-ES, see Paolella, Fundamental Statistics, 2018, §4.4; and the Wikipedia entry; both of which contain Matlab codes. The latter also discusses recent extensions of the baseline construct.

Proposition (Fitzpatrick, 13.15, The Mean Value Proposition): Let $\mathbf{x}$ be a point in $\mathbb{R}^n$ and let r be a positive number. Suppose that the function $f : \mathcal{B}_r(\mathbf{x}) \to \mathbb{R}$ has first-order partial derivatives. Then if the point $\mathbf{x}+\mathbf{h}$ belongs to $\mathcal{B}_r(\mathbf{x})$, there are points $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n \in \mathcal{B}_r(\mathbf{x})$ such that

$$f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) = \sum_{i=1}^{n} h_i \frac{\partial f}{\partial x_i}(\mathbf{z}_i), \tag{4.70}$$

For $h = (0, \ldots, a, 0, \ldots, 0)$, with value a in the ith position, this is (4.69), with $z_i = x + \theta a \mathbf{e}_i = (x_1, x_2, \ldots, x_i + \theta a, x_{i+1}, \ldots, x_n)$.

and

Note with my h and z_i just given, and recalling $0 < \theta < 1$, $\|x - z_i\| = \theta|a| < |a| = \|h\|$, in agreement with:

$$\|\mathbf{x} - \mathbf{z}_i\| < \|\mathbf{h}\| \quad \text{for each index } i \text{ with } 1 \le i \le n.$$

Proof: We prove the result with $n = 3$. From this, it will be clear that the general result is also true. The trick is to expand the difference $f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x})$. We have

$$\begin{aligned} f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) =& f(x_1+h_1, x_2+h_2, x_3+h_3) - f(x_1, x_2, x_3) \\ =& f(x_1+h_1, x_2+h_2, x_3+h_3) - f(x_1+h_1, x_2+h_2, x_3) \\ &+ f(x_1+h_1, x_2+h_2, x_3) - f(x_1+h_1, x_2, x_3) \\ &+ f(x_1+h_1, x_2, x_3) - f(x_1, x_2, x_3). \end{aligned} \tag{4.71}$$

The 2nd (3rd, 4th) line in (4.71) shows changes in the 3rd (2nd, 1st) component, respectively.

We apply the previous Mean Value Lemma (Fitzpatrick, 13.14) to each of these differences to find numbers θ_1, θ_2, and θ_3 in the open interval $(0, 1)$ with

$$\begin{aligned} f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) =& \frac{\partial f}{\partial x_3}(x_1+h_1, x_2+h_2, x_3+\theta_3 h_3)\, h_3 \\ &+ \frac{\partial f}{\partial x_2}(x_1+h_1, x_2+\theta_2 h_2, x_3)\, h_2 \\ &+ \frac{\partial f}{\partial x_1}(x_1+\theta_1 h_1, x_2, x_3)\, h_1. \end{aligned} \tag{4.72}$$

In (4.72), use the substitutions

$$\begin{aligned} \mathbf{z}_1 &= (x_1+\theta_1 h_1, x_2, x_3), \\ \mathbf{z}_2 &= (x_1+h_1, x_2+\theta_2 h_2, x_3), \\ \mathbf{z}_3 &= (x_1+h_1, x_2+h_2, x_3+\theta_3 h_3), \end{aligned}$$

the result follows. For the above norm inequality, recall $\theta_1, \theta_2, \theta_3 \in (0, 1)$, so that

$$\begin{aligned} \|\mathbf{x} - \mathbf{z}_1\| &= \|\mathbf{z}_1 - \mathbf{x}\| = \|(\theta_1 h_1,\, 0,\, 0)\| = \theta_1|h_1| < \|\mathbf{h}\|, \\ \|\mathbf{x} - \mathbf{z}_2\| &= \|\mathbf{z}_2 - \mathbf{x}\| = \|(h_1,\, \theta_2 h_2,\, 0)\| < \|\mathbf{h}\|, \\ \|\mathbf{x} - \mathbf{z}_3\| &= \|\mathbf{z}_3 - \mathbf{x}\| = \|(h_1,\, h_2,\, \theta_3 h_3)\| < \|\mathbf{h}\|. \end{aligned}$$

We now turn to an analysis of the limit in (4.25) when the point $\mathbf{e}_i$ is replaced by a general nonzero point $\mathbf{p}$ in $\mathbb{R}^n$.

Definition: Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}$. Consider a function $f : \mathcal{O} \to \mathbb{R}$ and a nonzero point $\mathbf{p}$ in $\mathbb{R}^n$. If the limit exists, we define

$$(D_{\mathbf{p}} f)(\mathbf{x}) = \frac{\partial f}{\partial \mathbf{p}}(\mathbf{x}) = \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{p}) - f(\mathbf{x})}{t} \tag{4.73}$$

to be the directional derivative of the function $f : \mathcal{O} \to \mathbb{R}$ in the direction $\mathbf{p}$ at the point $\mathbf{x}$. Observe that, if $\mathbf{p} = \mathbf{e}_i$, then (4.73) is equivalent to $(D_i f)(\mathbf{x})$. Figure 34 illustrates this for a bivariate function at point $\mathbf{x} = (x_0, y_0)$ in the direction of vector $\mathbf{u} = (a, b)$.

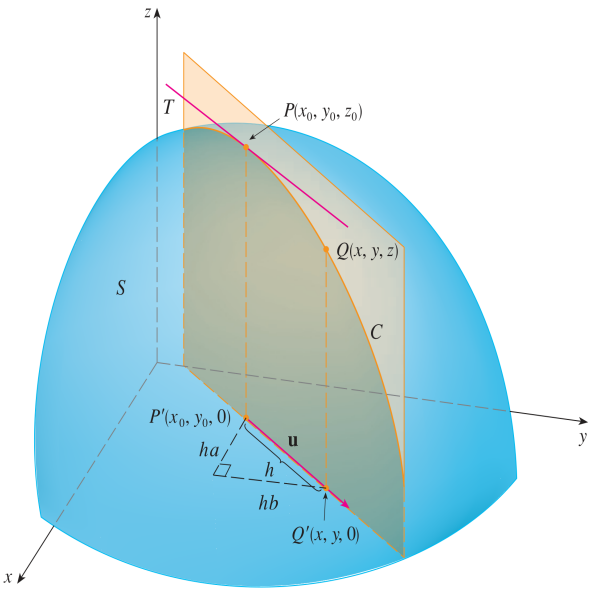


Figure 34: The slope of the tangent line T to slice C at the point P is the rate of change of $z = f(x, y)$ in the direction of vector $\mathbf{u} = (a, b)$, with $h = 1$ corresponding to $\|\mathbf{u}\| = 1$. From Stewart, Multivariate Calculus, 8th ed., 2016, p. 987.

Example 4.16 *(DePree and Swartz, Introduction to Real Analysis, 1988, p. 102) Even if a function has directional derivatives in all directions at a point, it may still fail to be continuous there. Consider the function*

$$f(x, y) = \begin{cases} xy^2 / \left(x^2 + y^4\right), & (x, y) \neq (0, 0), \\ 0, & (x, y) = (0, 0). \end{cases}$$

For $\mathbf{p} = (a, b)$ *with* $a \neq 0$,

$$\lim_{t \to 0} \frac{f(\mathbf{0} + t\mathbf{p}) - f(\mathbf{0})}{t} = \lim_{t \to 0} \frac{ab^2}{a^2 + t^2 b^4} = \frac{b^2}{a} = (D_{\mathbf{p}} f)(\mathbf{0});$$

if $a = 0$, $D_{\mathbf{p}} f(\mathbf{0}) = 0$. *Thus* f *has directional derivatives in all directions at* $\mathbf{0}$. *However,* f *is not continuous there, for on the curve* $y^2 = x$, f *has the value* $1/2$. ■

The terminology *directional derivative* is standard but is somewhat misleading, because the directional derivative depends not only on the direction of $\mathbf{p}$, but also on its length. To see this, recall (4.26): $(D_{\mathbf{p}}f)(\mathbf{x})$ is the derivative of $\gamma(t) = f(\mathbf{x}+t\mathbf{p})$ at 0. If we allow $\|\mathbf{p}\| \neq 1$, we can define $\phi(t) = f(\mathbf{x} + t\mathbf{p}/\|\mathbf{p}\|)$. Then

$$\gamma'(t) = \frac{d}{dt}f(\mathbf{x}+t\mathbf{p}) = \frac{d}{dt}f\left(\mathbf{x} + (t\|\mathbf{p}\|)\frac{\mathbf{p}}{\|\mathbf{p}\|}\right) = \|\mathbf{p}\|\phi'(t),$$

which results in different directional derivatives along the same direction. This might be undesirable.

Theorem (Fitzpatrick, 13.16, The Directional Derivative Theorem): Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ and suppose that the function $f : \mathcal{O} \to \mathbb{R}$ is continuously differentiable. Then for each point $\mathbf{x}$ in $\mathcal{O}$ and each nonzero point $\mathbf{p}$ in $\mathbb{R}^n$, the function $f : \mathcal{O} \to \mathbb{R}$ has a directional derivative in the direction $\mathbf{p}$ at the point $\mathbf{x}$ that is given by the formula

$$\frac{\partial f}{\partial \mathbf{p}}(\mathbf{x}) = \sum_{i=1}^{n} p_i \frac{\partial f}{\partial x_i}(\mathbf{x}), \qquad \mathbf{p} = (p_1, \ldots, p_n). \tag{4.74}$$

Proof: Since $\mathcal{O}$ is an open subset of $\mathbb{R}^n$, we can choose a positive number r such that the open ball $\mathcal{B}_r(\mathbf{x})$ is contained in $\mathcal{O}$. Then from the previous Mean Value Proposition (Fitzpatrick, 13.15), we see that, if t is any number with $|t|\|\mathbf{p}\| < r$, then there are n points $\mathbf{z}_1, \ldots, \mathbf{z}_n \in \mathcal{B}_r(\mathbf{x})$ such that

Recall in the statement of Prop. 13.15, "if the point $\mathbf{x} + \mathbf{h}$ belongs to $\mathcal{B}_r(\mathbf{x})$". Thus, $\mathbf{x} + t\mathbf{p} \in \mathcal{B}_r(\mathbf{x})$. Also, perhaps obviously, $|t|\|\mathbf{p}\|$ has the same direction as $\mathbf{p}$, for $t \neq 0$.

$$f(\mathbf{x}+t\mathbf{p}) - f(\mathbf{x}) = \sum_{i=1}^{n} tp_i \frac{\partial f}{\partial x_i}(\mathbf{z}_i) \tag{4.75}$$

and

$$\|\mathbf{z}_i - \mathbf{x}\| \leq |t|\|\mathbf{p}\| \quad \text{for each index } i \text{ with } 1 \leq i \leq n. \tag{4.76}$$

We can rewrite (4.75) as

$$\frac{f(\mathbf{x}+t\mathbf{p}) - f(\mathbf{x})}{t} = \sum_{i=1}^{n} p_i \frac{\partial f}{\partial x_i}(\mathbf{z}_i) \qquad \text{for } t \neq 0. \tag{4.77}$$

Since $\partial f/\partial x_i : \mathcal{O} \to \mathbb{R}$ is continuous at the point $\mathbf{x}$ for each index i with $1 \leq i \leq n$, it follows from (4.76) and (4.77) that and the definition of directional derivative

$$\lim_{t\to 0} \frac{f(\mathbf{x}+t\mathbf{p}) - f(\mathbf{x})}{t} = \lim_{t\to 0} \sum_{i=1}^{n} p_i \frac{\partial f}{\partial x_i}(\mathbf{z}_i) = \sum_{i=1}^{n} p_i \frac{\partial f}{\partial x_i}(\mathbf{x}).$$

This proves formula (4.74).

In view of formula (4.74), we introduce the following definition.

Definition: Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}$ and suppose that the function $f : \mathcal{O} \to \mathbb{R}$ has first-order partial derivatives at $\mathbf{x}$. (Not necessarily continuous.) We define the gradient of the function $f : \mathcal{O} \to \mathbb{R}$ at the point $\mathbf{x}$, denoted by $\nabla f(\mathbf{x})$, to be

the point in $\mathbb{R}^n$ given by and as previously defined in (4.22)

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}(\mathbf{x}), \frac{\partial f}{\partial x_2}(\mathbf{x}), \ldots, \frac{\partial f}{\partial x_n}(\mathbf{x}) \right). \tag{4.78}$$

Through the identification of points in $\mathbb{R}^n$ with vectors, $\nabla f(\mathbf{x})$ is often referred to as the gradient vector or derivative vector. Using the scalar product and the gradient, formula (4.74) can be compactly written as (Or as a regular matrix (here, vector) product, with the convention that ∇f is a row vector, and $\mathbf{p}$ is a column vector.)

$$\frac{d}{dt}[f(\mathbf{x}+t\mathbf{p})]\Big|_{t=0} = \frac{\partial f}{\partial \mathbf{p}}(\mathbf{x}) = \langle \nabla f(\mathbf{x}), \mathbf{p} \rangle. \tag{4.79}$$

It is also useful to observe a slight extension of (4.79): Replacing the point $\mathbf{x}$ in the latter two quantities; and not the first! with the point $\mathbf{x}+t\mathbf{p}$, it follows that and note that the left hand sides of (4.79) and (4.80) are not the same![30]

$$\frac{d}{dt}[f(\mathbf{x}+t\mathbf{p})] = \langle \nabla f(\mathbf{x}+t\mathbf{p}), \mathbf{p} \rangle, \tag{4.80}$$

provided that the segment between $\mathbf{x}$ and $\mathbf{x}+t\mathbf{p}$ lies in $\mathcal{O}$.

For further clarity, let $\phi(t; \mathbf{x}, \mathbf{p}) = \phi(t) = f(\mathbf{x}+t\mathbf{p})$, so that, from definition (4.73),

$$\phi'(t) = \frac{d}{dt} f(\mathbf{x}+t\mathbf{p}) = \frac{\partial f}{\partial \mathbf{p}}(\mathbf{x}+t\mathbf{p}),$$

and (4.79) is

$$\phi'(0) = \frac{d}{dt}[f(\mathbf{x}+t\mathbf{p})]\Big|_{t=0} = \frac{\partial f}{\partial \mathbf{p}}(\mathbf{x}).$$

Summarizing, from result (4.74) and using notation from (4.78),

$$\phi'(t) = \frac{\partial f}{\partial \mathbf{p}}(\mathbf{x}+t\mathbf{p}) = \langle \nabla f(\mathbf{x}+t\mathbf{p}), \mathbf{p} \rangle, \qquad \phi'(0) = \frac{\partial f}{\partial \mathbf{p}}(\mathbf{x}) = \langle \nabla f(\mathbf{x}), \mathbf{p} \rangle.$$

Theorem (Fitzpatrick, 13.17, The Mean Value Theorem): Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ and suppose that the function $f : \mathcal{O} \to \mathbb{R}$ is continuously differentiable. Because we need the Directional Derivative Theorem. If the segment joining the points $\mathbf{x}$ and $\mathbf{x}+\mathbf{h}$ lies in $\mathcal{O}$, then there is a number θ with $0 < \theta < 1$ such that $\phi(1) - \phi(0) =$

$$f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) = \langle \nabla f(\mathbf{x}+\theta\mathbf{h}), \mathbf{h} \rangle. \tag{4.81}$$

Proof: Since $\mathcal{O}$ is an open subset of $\mathbb{R}^n$, we can select an open interval of real numbers I, which contains the numbers 0 and 1, such that $\mathbf{x}+t\mathbf{h}$ belongs to $\mathcal{O}$ for each t in I. Define

$$\phi(t) = f(\mathbf{x}+t\mathbf{h}) \quad \text{for each } t \text{ in } I.$$

[30] In (4.80), Fitzpatrick also included that $0 \le t \le 1$, which does not seem correct. His subsequent statement "provided that the segment between $\mathbf{x}$ and $\mathbf{x}+t\mathbf{p}$ lies in $\mathcal{O}$" is what is required.

Using the slight generalization of the Directional Derivative Theorem stated as formula (4.80), we see that

$$\phi'(t) = \langle \nabla f(\mathbf{x} + t\mathbf{h}), \mathbf{h} \rangle \quad \text{for each } t \text{ in } I. \tag{4.82}$$

Thus, we can apply the Mean Value Theorem for functions of a single real variable to the restriction of the function $\phi : I \to \mathbb{R}$ to the closed interval $[0, 1]$ in order to select a number θ with $0 < \theta < 1$ such that

$$\phi(1) - \phi(0) = \phi'(\theta).$$

Using (4.82) and the definition of $\phi : [0, 1] \to \mathbb{R}$, it is clear that this formula is a restatement of (4.81).

In the case where $\mathbf{p}$ is a point in $\mathbb{R}^n$ of norm 1, a directional derivative in the direction $\mathbf{p}$ can be interpreted as a rate of change. To see this, let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}$ and suppose that the function $f : \mathcal{O} \to \mathbb{R}$ is continuously differentiable. Then if the point $\mathbf{p}$ is of norm 1 and t is a positive real number,

$$t = \|t\mathbf{p}\|$$

so, if t is positive and sufficiently small, so as to ensure that $x + t\mathbf{p}$ is in $\mathcal{O}$,

$$\frac{f(\mathbf{x} + t\mathbf{p}) - f(\mathbf{x})}{t} = \frac{f(\mathbf{x} + t\mathbf{p}) - f(\mathbf{x})}{\|t\mathbf{p}\|}.$$

In view of this, if the norm of $\mathbf{p}$ is 1, then it is reasonable to call $\partial f / \partial \mathbf{p}(\mathbf{x})$ as defined in (4.73) the rate of change of the function $f : \mathcal{O} \to \mathbb{R}$ in the direction $\mathbf{p}$ at the point $\mathbf{x}$.

Corollary (Fitzpatrick, 13.18): Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}$ and suppose that the function $f : \mathcal{O} \to \mathbb{R}$ is continuously differentiable. If $\nabla f(\mathbf{x}) \neq \mathbf{0}$, then the direction of norm 1 at the point $\mathbf{x}$ in which the function $f : \mathcal{O} \to \mathbb{R}$ is increasing the fastest is the direction $\mathbf{p}_0$ defined by

$$\mathbf{p}_0 = \frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|}. \tag{4.83}$$

Proof: Using formula (4.79) and the Cauchy-Schwarz inequality (1.9) or (3.20), it follows that, if $\mathbf{p}$ is any point in $\mathbb{R}^n$ of norm 1, then

$$\left| \frac{\partial f}{\partial \mathbf{p}}(\mathbf{x}) \right| = |\langle \nabla f(\mathbf{x}), \mathbf{p} \rangle| \leq \|\nabla f(\mathbf{x})\| \cdot \|\mathbf{p}\| = \|\nabla f(\mathbf{x})\|. \tag{4.84}$$

On the other hand, if $\mathbf{p}_0$ is defined by (4.83), then $\mathbf{p}_0$ has norm 1, and using (4.79), it follows that

$$\frac{\partial f}{\partial \mathbf{p}_0}(\mathbf{x}) = \langle \nabla f(\mathbf{x}), \mathbf{p}_0 \rangle = \left\langle \nabla f(\mathbf{x}), \frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|} \right\rangle = \|\nabla f(\mathbf{x})\|.$$

This calculation, together with inequality (4.84), implies that, if $\mathbf{p}$ has norm 1, then

$$\frac{\partial f}{\partial \mathbf{p}}(\mathbf{x}) \leq \frac{\partial f}{\partial \mathbf{p}_0}(\mathbf{x}).$$

Example 4.17 *Define*

$$f(x,y) = e^{x^2-y^2} \quad \text{for } (x,y) \in \mathbb{R}^2.$$

The function $f : \mathbb{R}^2 \to \mathbb{R}$ is continuously differentiable. A short calculation shows that

$$\frac{\partial f}{\partial x}(1,1) = 2 \quad \text{and} \quad \frac{\partial f}{\partial y}(1,1) = -2.$$

Thus, $\nabla f(1,1) = (2,-2)$, so the direction in which the function $f : \mathbb{R}^2 \to \mathbb{R}$ is increasing the fastest at the point $(1,1)$ is given by the vector $(1/\sqrt{2}, -1/\sqrt{2})$. ■

We are finally in a position to provide a definitive proof that $f \in \mathcal{C}^1(\mathcal{O}) \Rightarrow f \in \mathcal{C}^0(\mathcal{O})$, for general n, as stated in (4.46).

Theorem (Fitzpatrick, 13.20): Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ and suppose that the function $f : \mathcal{O} \to \mathbb{R}$ is continuously differentiable, i.e., $f \in \mathcal{C}^1(\mathcal{O})$. Then f is continuous.

Proof: Let $\mathbf{x}$ be a point in $\mathcal{O}$. We need to show that the function $f : \mathcal{O} \to \mathbb{R}$ is continuous at $\mathbf{x}$. We directly apply the sequential definition of continuity (4.17). First, since $\mathbf{x}$ is an interior point of $\mathcal{O}$, we can select a positive number r such that the open ball $\mathcal{B}_r(\mathbf{x})$ is contained in $\mathcal{O}$. Let $\{\mathbf{x}_k\}$ be a sequence in $\mathcal{B}_r(\mathbf{x})$ that converges to $\mathbf{x}$. For each natural number k, set $\mathbf{h}_k = \mathbf{x}_k - \mathbf{x}$ and apply the Mean Value Theorem to select a number θ_k with $0 < \theta_k < 1$ such that

$$f(\mathbf{x}_k) - f(\mathbf{x}) = f(\mathbf{x} + \mathbf{h}_k) - f(\mathbf{x}) = \langle \nabla f(\mathbf{x} + \theta_k \mathbf{h}_k), \mathbf{h}_k \rangle . \tag{4.85}$$

Now observe that

$$\lim_{k\to\infty} \mathbf{h}_k = \mathbf{0} \quad \text{and} \quad \lim_{k\to\infty} [\mathbf{x} + \theta_k \mathbf{h}_k] = \mathbf{x}.$$

Since $f : \mathcal{O} \to \mathbb{R}$ is continuously differentiable, it follows that (Each component of the gradient, and thus the entire vector itself, is continuous.)

$$\lim_{k\to\infty} \nabla f(\mathbf{x} + \theta_k \mathbf{h}_k) = \nabla f(\mathbf{x}).$$

Thus, since (4.85) holds for every index k, we conclude that

$$\lim_{k\to\infty} [f(\mathbf{x}_k) - f(\mathbf{x})] = \langle \nabla f(\mathbf{x}), \mathbf{0} \rangle = 0,$$

which means that the image sequence $\{f(\mathbf{x}_k)\}$ converges to $f(\mathbf{x})$.

Corollary: Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ and suppose that the function $f : \mathcal{O} \to \mathbb{R}$ has continuous second-order partial derivatives. Then the function $f : \mathcal{O} \to \mathbb{R}$ is continuously differentiable.

Proof: For each index i with $1 \le i \le n$, the function $\partial f / \partial x_i : \mathcal{O} \to \mathbb{R}$ is continuously differentiable, and hence, by Theorem (Fitzpatrick, 13.20), it is continuous. This is precisely what it means for the function $f : \mathbb{R}^n \to \mathbb{R}$ to be continuously differentiable.

4.6 The Jacobian and the Chain Rule

We now turn to multivariate functions of the form $\mathbf{f} : A \subset \mathbb{R}^n \to \mathbb{R}^m$, and note that we use bold face for the function name in order to indicate that $m > 1$.[31]

4.6.1 The Jacobian

Consider a multivariate function $\mathbf{f} : A \subset \mathbb{R}^n \to \mathbb{R}^m$ with A an open set, where $\mathbf{f}$ is such that $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))$, $f_i : A \to \mathbb{R}$, $i = 1, \ldots, m$, for all $\mathbf{x} = (x_1, \ldots, x_n)' \in A$, recalling the component notation from (4.7) and (4.8). If each partial derivative, $\partial f_i(\mathbf{x}_0)/\partial x_j$, $i = 1, \ldots, m$, $j = 1, \ldots, n$, exists, then the *total derivative* of $\mathbf{f}$ at $\mathbf{x}_0 \in A$ is the $m \times n$ matrix

$$\mathbf{f}'(\mathbf{x}_0) := \mathbf{J}_{\mathbf{f}}(\mathbf{x}_0) := \begin{pmatrix} \dfrac{\partial f_1}{\partial x_1}(\mathbf{x}_0) & \cdots & \dfrac{\partial f_1}{\partial x_n}(\mathbf{x}_0) \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_m}{\partial x_1}(\mathbf{x}_0) & \cdots & \dfrac{\partial f_m}{\partial x_n}(\mathbf{x}_0) \end{pmatrix} = \begin{pmatrix} (\operatorname{grad} f_1)(\mathbf{x}_0) \\ \vdots \\ (\operatorname{grad} f_m)(\mathbf{x}_0) \end{pmatrix}, \tag{4.86}$$

also referred to as the *Jacobian matrix* of $\mathbf{f}$ at $\mathbf{x}_0$.[32] When $m = 1$, the total derivative is just the gradient (4.22). Analogous to the $m = 1$ case from (4.34), let

$$\mathbf{T}(\mathbf{x}) = \mathbf{f}(\mathbf{x}_0) + \mathbf{J}_{\mathbf{f}}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) := \mathbf{f}(\mathbf{x}_0) + \mathbf{Df}(\mathbf{x}_0, \mathbf{x} - \mathbf{x}_0), \tag{4.87}$$

where $\mathbf{Df}(\mathbf{x}_0, \mathbf{h})$, defined on the rhs, is the total differential, analogous to the $m = 1$ case in (4.35), $\mathrm{d}f(\mathbf{x}, \mathbf{h}) = (\operatorname{grad} f)(\mathbf{x}) \cdot \mathbf{h}$.

Warning: Some authors (such as Petrovic; see below in equation (4.98); and Fitzpatrick; see (4.159)) refer to $\mathbf{J}_{\mathbf{f}}(\mathbf{x}_0)$ as $\mathbf{Df}(\mathbf{x}_0)$, which conflicts with our usage of $\mathbf{Df}$ in (4.87). Notice, at least, that the total differential $\mathbf{Df}(\mathbf{x}_0, \mathbf{h})$ takes two arguments, whereas the Jacobian $\mathbf{J}_{\mathbf{f}}(\mathbf{x}_0)$, also called the **total derivative** (Petrovic), or **derivative matrix** (Fitzpatrick), and its common equivalent notation $\mathbf{Df}(\mathbf{x}_0)$, take only one argument, and thus are distinguishable in context.

One resolution is, within this document, to convert all occurrences of the latter to the $\mathbf{J}_{\mathbf{f}}$ notation (or vice-versa). We opted not to do this, because both notations are popular in the literature, and it is best the reader becomes aware of this.

[31] The notes in this subsection, and §4.7, were assembled from me (for my math appendix in Fundamental Probability) over 20 years ago, and came from a compilation of several books. These included Trench (2003), the most recent version of which is Trench (Introduction to Real Analysis, Free Hyperlinked Edition 2.04, December, 2013); Lang (Undergraduate Analysis, 2nd ed., 1997); Hubbard and Hubbard (Vector Calculus, Linear Algebra, and Differential Forms: A Unified Approach, now in its 5th edition, 2015), and Protter and Morrey (A First Course in Real Analysis, 2nd edition, 1991).

[32] After the prolific Carl Gustav Jacob Jacobi (1804–1851), who made contributions in several branches of mathematics, including the study of functional determinants. Though the theory goes back (at least) to Cauchy in 1815, Jacobi's 1841 memoir *De determinantibus functionalibus* had the first modern definition of determinant, and the first use of the word Jacobian was by Sylvester in 1853. Jacobi is also remembered as an excellent teacher who introduced the "seminar method" for teaching the latest advances in math (whereby students present and discuss current articles and papers.)

Let $\mathbf{f} = (f_1, f_2, \ldots, f_m) : A \subset \mathbb{R}^n \to \mathbb{R}^m$, with A open, and let $\mathbf{x}_0 \in A$. We state two equivalent definitions of differentiability. The first is: Function $\mathbf{f}$ is said to be *differentiable at* $\mathbf{x}_0 \in A$ if each f_i, $i = 1, \ldots, m$, is differentiable at $\mathbf{x}_0 \in A$. Here is the second:

Differentiability of $\mathbf{f}$ is defined analogously to the $m = 1$ case above, namely that an $m \times n$ matrix, called the (total) derivative of $\mathbf{f}$ at $\mathbf{x}_0$ exists, and the analogies of (4.31) and (4.32) hold: There exists a *tangent map* $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^m$ of $\mathbf{f}$ at $\mathbf{x}_0$, such that

$$\mathbf{f}(\mathbf{x}_0) = \mathbf{T}(\mathbf{x}_0) \quad \text{and} \quad \lim_{\mathbf{x}\to\mathbf{x}_0} \frac{\mathbf{f}(\mathbf{x}) - \mathbf{T}(\mathbf{x})}{\|\mathbf{x} - \mathbf{x}_0\|} = 0, \tag{4.88}$$

where $\mathbf{T}(\mathbf{x})$ is given in (4.87). As also stated for the $m = 1$ case above just after (4.32), the actual definition does not include this latter specification of $\mathbf{T}(\mathbf{x})$, but rather only the existence of a tangent map and "a" total derivative matrix. The next theorem shows that this specification must in fact be (4.87) using the Jacobian matrix (4.86).

Theorem: $\mathbf{f}$ *is differentiable at* $\mathbf{x}_0 \in A$ *iff the Jacobian* $\mathbf{J}_\mathbf{f}(\mathbf{x}_0)$ *exists,* $\mathbf{f}(\mathbf{x}_0) = \mathbf{T}(\mathbf{x}_0)$*, and*

$$\lim_{\mathbf{x}\to\mathbf{x}_0} \frac{\|\mathbf{f}(\mathbf{x}) - \mathbf{T}(\mathbf{x})\|}{\|\mathbf{x} - \mathbf{x}_0\|} = 0. \tag{4.89}$$

Proof:

($\Rightarrow$) Assume $\mathbf{f}$ is differentiable at $\mathbf{x}_0 \in A$. Then, by definition, $(\text{grad } f_i)(\mathbf{x}_0)$ exists, $i = 1, \ldots, m$, so that the form of (4.86) shows that $\mathbf{J}_\mathbf{f}(\mathbf{x}_0)$ exists. Next, differentiability of $\mathbf{f}$ means that there exists a tangent map of each f_i at $\mathbf{x}_0$, given by, say, $T_i(\mathbf{x}) = f_i(\mathbf{x}_0) + (\text{grad } f_i)(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$, and, for each i, $f_i(\mathbf{x}_0) = T_i(\mathbf{x}_0)$. Thus, taking

$$\begin{aligned} \mathbf{T}(\mathbf{x}) = \begin{bmatrix} T_1(\mathbf{x}) \\ \vdots \\ T_m(\mathbf{x}) \end{bmatrix} &= \begin{bmatrix} f_1(\mathbf{x}_0) + (\text{grad } f_1)(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) \\ \vdots \\ f_m(\mathbf{x}_0) + (\text{grad } f_m)(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) \end{bmatrix} \\ &= \mathbf{f}(\mathbf{x}_0) + \mathbf{J}_\mathbf{f}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0), \end{aligned} \tag{4.90}$$

it is clear that $\mathbf{T}(\mathbf{x}_0) = (f_1(\mathbf{x}_0), \ldots, f_m(\mathbf{x}_0))' = \mathbf{f}(\mathbf{x}_0)$. Lastly, from (4.31), differentiability of $\mathbf{f}$ implies that, for each $i = 1, \ldots, m$,

$$\lim_{\mathbf{x}\to\mathbf{x}_0} \frac{f_i(\mathbf{x}) - T_i(\mathbf{x})}{\|\mathbf{x} - \mathbf{x}_0\|} = 0 = \lim_{\mathbf{x}\to\mathbf{x}_0} \frac{|f_i(\mathbf{x}) - T_i(\mathbf{x})|}{\|\mathbf{x} - \mathbf{x}_0\|}. \tag{4.91}$$

Next, note that, from a repeated application of the triangle inequality, for any real vector $\mathbf{z} = (z_1, \ldots, z_m)$, $\|\mathbf{z}\| \le |z_1| + \cdots + |z_m|$. Thus, with

$$\mathbf{z} = \frac{\mathbf{f}(\mathbf{x}) - \mathbf{T}(\mathbf{x})}{\|\mathbf{x} - \mathbf{x}_0\|} = \frac{1}{\|\mathbf{x} - \mathbf{x}_0\|} \begin{bmatrix} f_1(\mathbf{x}) - T_1(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) - T_m(\mathbf{x}) \end{bmatrix},$$

it follows that

$$\|\mathbf{z}\| = \frac{\|\mathbf{f}(\mathbf{x}) - \mathbf{T}(\mathbf{x})\|}{\|\mathbf{x} - \mathbf{x}_0\|} \le \sum_{i=1}^{m} \frac{|f_i(\mathbf{x}) - T_i(\mathbf{x})|}{\|\mathbf{x} - \mathbf{x}_0\|},$$

i.e., (4.89) follows from (4.91).

($\Leftarrow$) If $\mathbf{J_f}(\mathbf{x}_0)$ exists, then $(\operatorname{grad} f_i)(\mathbf{x}_0)$ exists, $i = 1, \ldots, m$. From (4.90), if $\mathbf{T}(\mathbf{x}_0) = \mathbf{f}(\mathbf{x}_0)$, then $f_i(\mathbf{x}_0) = T_i(\mathbf{x}_0)$, $i = 1, \ldots, m$. Lastly, it trivially follows from the definition of the norm that

$$\frac{|f_i(\mathbf{x}) - T_i(\mathbf{x})|}{\|\mathbf{x} - \mathbf{x}_0\|} \leq \frac{\|\mathbf{f}(\mathbf{x}) - \mathbf{T}(\mathbf{x})\|}{\|\mathbf{x} - \mathbf{x}_0\|},$$

so that (4.91) follows from (4.89).

Paralleling the $m = 1$ case from §4.3, $\mathbf{f}$ is *differentiable on* A if $\mathbf{f}$ is differentiable at each $\mathbf{x}_0 \in A$. Furthermore, if $\mathbf{f}$ is differentiable and all the partial derivatives of each f_i are continuous, then $\mathbf{f}$ is *continuously differentiable*, and we write $\mathbf{f} \in \mathcal{C}^1(A^o)$, where A^o is the interior of the domain A of $\mathbf{f}$.

Example 4.18 *Let* $\mathbf{f} : \mathbb{R}^2 \to \mathbb{R}^3, (x, y) \mapsto (ye^x, x^2y^3, -x)$. *Then, from* (4.86),

$$\mathbf{f}'(x, y) = \mathbf{J_f}(x, y) = \begin{pmatrix} ye^x & e^x \\ 2xy^3 & x^2 3y^2 \\ -1 & 0 \end{pmatrix},$$

and $\mathbf{f}$ *is continuously differentiable.* ■

In §2.2.1, we showed (the easy, standard proof) that, for $f : A \subset \mathbb{R} \to \mathbb{R}$, if f is differentiable at the point $a \in A$, then f is continuous at a. This was extended to $f : A \subset \mathbb{R}^n \to \mathbb{R}$ in (4.43). We now state and prove the result for $\mathbf{f} : A \subset \mathbb{R}^n \to \mathbb{R}^m$. This obviously subsumes the two mentioned special cases.

If $\mathbf{f}$ is differentiable at $\mathbf{x}_0 \in A$, then $\mathbf{f}$ is continuous at $\mathbf{x}_0$. (4.92)

The proof hinges on the two most important inequalities in analysis: From (4.89), $\exists\, \delta^* > 0$ such that, for $\mathbf{x} \in A$, if $\|\mathbf{x} - \mathbf{x}_0\| < \delta^*$, then $\|\mathbf{f}(\mathbf{x}) - \mathbf{T}(\mathbf{x})\| < \|\mathbf{x} - \mathbf{x}_0\|$, where $\mathbf{T}(\mathbf{x}) = \mathbf{f}(\mathbf{x}_0) + \mathbf{J_f}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$. Let $\mathbf{K} = \mathbf{J_f}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$ so that $\mathbf{T}(\mathbf{x}) = \mathbf{f}(\mathbf{x}_0) + \mathbf{K}$. If $\|\mathbf{x} - \mathbf{x}_0\| < \delta^*$, then, from the triangle inequality (1.10),

$$\begin{aligned} \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)\| &= \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0) - \mathbf{K} + \mathbf{K}\| = \|\mathbf{f}(\mathbf{x}) - \mathbf{T}(\mathbf{x}) + \mathbf{K}\| \\ &\leq \|\mathbf{f}(\mathbf{x}) - \mathbf{T}(\mathbf{x})\| + \|\mathbf{K}\| < \|\mathbf{x} - \mathbf{x}_0\| + \|\mathbf{K}\|. \end{aligned} \tag{4.93}$$

From (4.90), with row vector $\mathbf{w}_i = (w_{i1}, \ldots, w_{in}) := (\operatorname{grad} f_i)(\mathbf{x}_0)$ and column vector $\mathbf{z}_i = (z_{i1}, \ldots, z_{in}) := (\mathbf{x} - \mathbf{x}_0)$,

$$\|\mathbf{K}\|^2 = \sum_{i=1}^{m} \left[(\operatorname{grad} f_i)(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) \right]^2 = \sum_{i=1}^{m} \left[\mathbf{w}_i \mathbf{z}_i\right]^2 = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} w_{ij} z_{ij} \right)^2.$$

For each $i = 1, \ldots, m$, the Cauchy-Schwarz inequality (1.9) implies

$$\left(\sum_{j=1}^{n} w_{ij} z_{ij} \right)^2 \leq \left(\sum_{j=1}^{n} w_{ij}^2 \right) \left(\sum_{j=1}^{n} z_{ij}^2 \right) = \|(\operatorname{grad} f_i)(\mathbf{x}_0)\|^2 \|(\mathbf{x} - \mathbf{x}_0)\|^2,$$

so that

$$\begin{aligned} \|\mathbf{K}\| &\leq \left(\sum_{i=1}^{m} \|(\operatorname{grad} f_i)(\mathbf{x}_0)\|^2 \|(\mathbf{x} - \mathbf{x}_0)\|^2 \right)^{1/2} \\ &= \|(\mathbf{x} - \mathbf{x}_0)\| \left(\sum_{i=1}^{m} \|(\operatorname{grad} f_i)(\mathbf{x}_0)\|^2 \right)^{1/2} =: \|(\mathbf{x} - \mathbf{x}_0)\| \, G. \end{aligned} \tag{4.94}$$

Thus, from (4.93) and (4.94),

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)\| < \|\mathbf{x} - \mathbf{x}_0\| + \|\mathbf{K}\| < \|\mathbf{x} - \mathbf{x}_0\|(1 + G). \tag{4.95}$$

Because we assume that $\mathbf{f}$ is differentiable at $\mathbf{x}_0$, G is finite. Thus, for a given $\epsilon > 0$, we can find a $\delta > 0$ such that, if $\|\mathbf{x} - \mathbf{x}_0\| < \delta$ and $\mathbf{x} \in A$, then $\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)\| < \epsilon$. In particular, from (4.95), $\delta = \min(\delta^*, \epsilon/(1 + G))$.

In (4.46) and page 229, we demonstrated that, if $f : \mathcal{O} \to \mathbb{R}$ is continuously differentiable on open domain $\mathcal{O}$, then f is continuous. That is, for $m = 1$, $f \in \mathcal{C}^1(\mathcal{O}) \Rightarrow \mathcal{C}^0(\mathcal{O})$. We now state the result for the case of multivariate function $\mathbf{f} : \mathcal{O} \subset \mathbb{R}^n \to \mathbb{R}^m$. See, e.g., Hubbard and Hubbard (2002, p. 159 and p. 680) for detailed proofs.

If all the partial derivatives of $\mathbf{f} = (f_1, \ldots, f_m) : \mathcal{O} \subset \mathbb{R}^n \to \mathbb{R}^m$ exist and are continuous on $\mathcal{O}$, then $\mathbf{f}$ is differentiable on $\mathcal{O}$ with derivative $\mathbf{J}_\mathbf{f}$.

4.6.2 The Chain Rule

The next result is the chain rule, providing a big generalization to the univariate case (2.40). We provide two proofs; one below, and another in §4.9. Before stating the chain rule, it is useful to review the definition of differentiability, as given either in (4.89); or below in (4.98) and (4.99).

Let $\mathbf{f} : A \subset \mathbb{R}^n \to \mathbb{R}^m$ and $\mathbf{g} : B \subset \mathbb{R}^m \to \mathbb{R}^p$ with A and B open sets and $\mathbf{f}(A) \subset B$. If $\mathbf{f}$ is differentiable at $\mathbf{x} \in A$ and $\mathbf{g}$ is differentiable at $\mathbf{f}(\mathbf{x})$, then the composite function $\mathbf{g} \circ \mathbf{f}$ is differentiable at $\mathbf{x}$, with derivative

$$\mathbf{J}_{\mathbf{g}\circ\mathbf{f}}(\mathbf{x}) = \mathbf{J}_\mathbf{g}(\mathbf{f}(\mathbf{x}))\,\mathbf{J}_\mathbf{f}(\mathbf{x}). \tag{4.96}$$

Example 4.19 *As in Example 4.18, let $\boldsymbol{f} : \mathbb{R}^2 \to \mathbb{R}^3, (x, y) \mapsto (ye^x, x^2y^3, -x)$, and also let $\mathbf{g} : (\mathbb{R}_{>0} \times \mathbb{R}) \to \mathbb{R}^2, (x, y) \mapsto (\ln x, x + 2y) = (g_1(x, y), g_2(x, y))$. The function $\mathbf{g}$ is continuously differentiable with derivative at $(x, y) \in (\mathbb{R}_{>0} \times \mathbb{R})$:*

$$\mathbf{g}'(x, y) = \mathbf{J}_\mathbf{g}(x, y) = \begin{bmatrix} 1/x & 0 \\ 1 & 2 \end{bmatrix}.$$

Let $\mathbf{h} = \mathbf{f} \circ \mathbf{g}$. The composition $\mathbf{h}$ is continuously differentiable and its derivative at $(x, y) \in (\mathbb{R}_{>0} \times \mathbb{R})$ is given by

$$\begin{aligned}
\mathbf{h}'(x, y) &= \mathbf{J}_\mathbf{h}(x, y) = \mathbf{J}_\mathbf{f}(\mathbf{g}(x, y)) \cdot \mathbf{J}_\mathbf{g}(x, y) \\
&= \begin{bmatrix} (x + 2y)x & x \\ 2\ln(x)(x + 2y)^3 & (\ln(x))^2 3(x + 2y)^2 \\ -1 & 0 \end{bmatrix} \cdot \begin{pmatrix} 1/x & 0 \\ 1 & 2 \end{pmatrix} \\
&= \begin{bmatrix} (x + 2y) + x & 2x \\ (1/x)2\ln x(x + 2y)^3 + (\ln x)^2 3(x + 2y)^2 & 2(\ln x)^2 3(x + 2y)^2 \\ -1/x & 0 \end{bmatrix}.
\end{aligned}$$

The reader is encouraged to calculate an expression for $\mathbf{h}(x, y)$ and compute $\mathbf{J}_\mathbf{h}$ directly. ■

For the chain rule (4.96), a special case of interest is $n = p = 1$, i.e., $\mathbf{f} = (f_1, \ldots, f_m) : A \subset \mathbb{R} \to \mathbb{R}^m$ and $g : B \subset \mathbb{R}^m \to \mathbb{R}$. Then for $x \in A$, $\mathbf{J}_g$ is a row vector and $\mathbf{J}_\mathbf{f}$ is a column vector, so that (4.96) simplifies to (see also below, (4.109) and (4.110), for more detail)

$$\mathbf{J}_{g\circ\mathbf{f}}(x) = \sum_{i=1}^{m} \frac{\partial g}{\partial f_i}(\mathbf{f}(x))\frac{df_i}{dx}(x), \tag{4.97}$$

where $\partial g/\partial f_i$ denotes the ith partial derivative of g.[33] A mnemonic version of this formula is, with $h = g \circ \mathbf{f}$,

$$\frac{dh}{dx} = \sum_{i=1}^{m} \frac{\partial h}{\partial f_i}\frac{df_i}{dx}.$$

Example 4.20 *Assume that the United States GDP, denoted by P, is a continuously differentiable function of the capital, C, and the work force, W. Moreover, assume C and W are continuously differentiable functions of time, t. Then P is a continuously differentiable function of t and economists would write:*

$$\frac{dP}{dt} = \frac{\partial P}{\partial C}\frac{dC}{dt} + \frac{\partial P}{\partial W}\frac{dW}{dt},$$

showing us how the change of P can be split into a part due to the decrease or increase of C and another due to the change of W. ■

The following example is useful, as it shows what can go wrong when at least one of the functions is not differentiable. In particular, there is no tangent map of function g at the point $(0, 0)$. Recall Example 4.12 in which the function has partial derivatives at each point in its domain, but at $(0, 0)$, there is no tangent map, and thus the function is not differentiable over its whole domain.

Example 4.21 *(Counterexample to Example 4.20) Consider the following functions:*

$$g : \mathbb{R}^2 \to \mathbb{R}, (x, y) \mapsto \begin{cases} \dfrac{x^3 + y^3}{x^2 + y^2}, & \text{if } (x, y) \neq (0, 0), \\ 0, & \text{if } (x, y) = (0, 0), \end{cases}$$

and $\mathbf{f} : \mathbb{R} \to \mathbb{R}^2$, given by $t \mapsto (t, t)$. The partial derivatives of g exist on the whole domain, in particular, at $(x, y) = (0, 0)$:

$$\frac{\partial g}{\partial x}(0, 0) = \lim_{x\to 0} \frac{g(x, 0) - g(0, 0)}{x - 0} = \lim_{x\to 0} \frac{x^3 + 0}{x(x^2 + 0)} = \lim_{x\to 0} 1 = 1 = \frac{\partial g}{\partial y}(0, 0),$$

but the partial derivatives of g are not continuous. However, $\mathbf{f}$ is continuously differentiable with derivative

$$\mathbf{f}' : \mathbb{R} \to \mathbb{R}^2, \quad t \mapsto \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

So, if the chain rule were applicable here, then the derivative of $h := g\circ\mathbf{f}$ at $t = 0$ is calculated to be

$$h'(0) = \mathbf{J}_g(\mathbf{f}(0))\mathbf{J}_\mathbf{f}(0) = (1 \quad 1)\begin{pmatrix} 1 \\ 1 \end{pmatrix} = 2.$$

[33] This notation is somewhat misleading, as f_i in $\partial g/\partial f_i$ is a function itself. Keep in mind that, in "$\partial g/\partial f_i$", the f_i could be replaced by, say, y_i, or any other name of variable as long as it can be easily inferred which partial derivative is meant.

On the other hand, we can calculate $h(t)$ for $t \in \mathbb{R}$ directly as

$$h(t) = \frac{t^3 + t^3}{t^2 + t^2} = \frac{2t^3}{2t^2} = t,$$

so that $h'(0) = 1$. This demonstrates that the chain rule generally does not hold when one of the functions is not continuously differentiable. ■

We include now an excerpt from Petrovic, pp. 355-6, repeating some of the above definitions, and including a proof of the Chain Rule (4.96), and an additional basic result. We will give yet another discussion of this topic in §4.9, and, there, yet another proof of the Chain Rule, which appears in (4.179). Recall the notational warning given above.

Definition: Let $\mathbf{f} = (f_1, f_2, \dots, f_m)$ be a function defined in an open ball $A \subset \mathbb{R}^n$, with values in $\mathbb{R}^m$, and let $\mathbf{a} \in A$. Then $\mathbf{f}$ is differentiable at $\mathbf{a}$ if and only if there exists an $m \times n$ matrix $\mathbf{Df}(\mathbf{a})$, called the (total) derivative of $\mathbf{f}$ at $\mathbf{a}$, such that

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{a}) + \mathbf{Df}(\mathbf{a})(\mathbf{x} - \mathbf{a}) + \mathbf{r}(\mathbf{x}), \quad \text{and} \tag{4.98}$$

$$\lim_{\mathbf{x} \to \mathbf{a}} \frac{\mathbf{r}(\mathbf{x})}{\|\mathbf{x} - \mathbf{a}\|} = \mathbf{0}. \tag{4.99}$$

If $\mathbf{f}$ is differentiable at every point of a set A, we say that it is differentiable on A.

We are making a standard identification between elements of the Euclidean space of dimension n, and $n \times 1$ matrices. For example, $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_m(\mathbf{x}))$ can be viewed as a column matrix of dimension $m \times 1$. That means that (4.98) states an equality between matrices. If we read it row by row, we can conclude several things. First, the rows of $\mathbf{Df}$ are precisely the partial derivatives of the functions $f_1, f_2, \dots, f_m$, so

$$\mathbf{Df}(\mathbf{a}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

Second, $\mathbf{f}$ is differentiable at $\mathbf{a}$ iff f_i is differentiable at $\mathbf{a}$ for all $1 \leq i \leq m$.

All the expected rules for derivatives hold, as shown next, simply because they hold for each of the component functions. Recall, e.g., (4.52), which states that, for $f, g : A \subset \mathbb{R}^n \to \mathbb{R}$ differentiable at $\mathbf{a} \in A^o$, $f \cdot g$ is differentiable at $\mathbf{a}$, with

$$\operatorname{grad}(f \cdot g)(\mathbf{a}) = g(\mathbf{a}) \operatorname{grad} f(\mathbf{a}) + f(\mathbf{a}) \operatorname{grad} g(\mathbf{a}), \tag{4.100}$$

which can be compared to part (d) of the next theorem.

Theorem: Let $\mathbf{f}, \mathbf{g} : A \subset \mathbb{R}^n \to \mathbb{R}^m$, A open, and let $\mathbf{a} \in A$. Also, let $\alpha \in \mathbb{R}$ and $\varphi : A \subset \mathbb{R}^n \to \mathbb{R}$. If $\mathbf{f}$, $\mathbf{g}$, and φ are differentiable at $\mathbf{a}$, then so are $\mathbf{f} + \mathbf{g}$, $\alpha\mathbf{f}$, $\varphi\mathbf{f}$, and $\mathbf{f} \cdot \mathbf{g}$:

(a) $\mathbf{D}(\alpha\mathbf{f})(\mathbf{a}) = \alpha\mathbf{Df}(\mathbf{a})$;

(b) $\mathbf{D}(\mathbf{f} + \mathbf{g})(\mathbf{a}) = \mathbf{Df}(\mathbf{a}) + \mathbf{Dg}(\mathbf{a})$;

(c) $\mathbf{D}(\varphi\mathbf{f})(\mathbf{a}) = \mathbf{f}(\mathbf{a})\mathbf{D}\varphi(\mathbf{a}) + \varphi(\mathbf{a})\mathbf{Df}(\mathbf{a})$;

(d) $\mathbf{D}(\mathbf{f} \cdot \mathbf{g})(\mathbf{a}) = \mathbf{g}(\mathbf{a})'\mathbf{Df}(\mathbf{a}) + \mathbf{f}(\mathbf{a})'\mathbf{Dg}(\mathbf{a})$.

Remarks:

1. The proofs of (a) and (b) follows easily from its $m = 1$ counterpart (4.50) and (4.51). Part (a) also follows from (c), as noted next.

2. Let $\varphi(\mathbf{a}) = \alpha$. Then $\mathbf{D}\varphi(\mathbf{a}) = \mathbf{0}$, of size $1 \times n$, and (c) implies (a).

3. The three objects in (c) must all be $m \times n$; and as $\mathbf{D}\varphi(\mathbf{a})$ is $1 \times n$, we require $\mathbf{f}(\mathbf{a})$ to be the column vector of size $m \times 1$. We apply the same to $\mathbf{g}(\mathbf{a})$.

4. For (d), note that $\mathbf{f} \cdot \mathbf{g}$ is a dot product, and thus a mapping from A to $\mathbb{R}$. That means $\mathbf{D}(\mathbf{f} \cdot \mathbf{g})(\mathbf{a})$ must be $1 \times n$. The rhs indeed has this dimension. We verify the rhs next.

5. Denote, as usual, the component functions of $\mathbf{f}$ as $(f_1, \ldots, f_m)$; and likewise for $\mathbf{g}$. To see how (d) follows from (4.100), first note that, as $\mathbf{D}$ preserves linearity, and from (4.100),

$$\begin{aligned}\mathbf{D}\left(\mathbf{f} \cdot \mathbf{g}\right)(\mathbf{a}) &= \mathbf{D}\left(\sum_{i=1}^{m} f_i\left(\mathbf{a}\right) g_i\left(\mathbf{a}\right)\right) = \sum_{i=1}^{m} \mathbf{D}\left(f_i\left(\mathbf{a}\right) g_i\left(\mathbf{a}\right)\right) \\ &= \sum_{i=1}^{m} \left[g_i\left(\mathbf{a}\right) \nabla f_i\left(\mathbf{a}\right) + f_i\left(\mathbf{a}\right) \nabla g_i\left(\mathbf{a}\right)\right]. \end{aligned} \tag{4.101}$$

   Next, (d) states (with each $\nabla f_i\left(\mathbf{a}\right)$ and $\nabla g_i\left(\mathbf{a}\right)$ being of size $1 \times n$, $i = 1, \ldots m$)

$$\begin{aligned}\mathbf{D}\left(\mathbf{f} \cdot \mathbf{g}\right)(\mathbf{a}) &= \left(g_1\left(\mathbf{a}\right), \ldots, g_m\left(\mathbf{a}\right)\right) \begin{bmatrix} \nabla f_1\left(\mathbf{a}\right) \\ \nabla f_2\left(\mathbf{a}\right) \\ \vdots \\ \nabla f_m\left(\mathbf{a}\right) \end{bmatrix} + \left(f_1\left(\mathbf{a}\right), \ldots, f_m\left(\mathbf{a}\right)\right) \begin{bmatrix} \nabla g_1\left(\mathbf{a}\right) \\ \nabla g_2\left(\mathbf{a}\right) \\ \vdots \\ \nabla g_m\left(\mathbf{a}\right) \end{bmatrix} \\ &= \left[g_1\left(\mathbf{a}\right) \nabla f_1\left(\mathbf{a}\right) + \cdots + g_m\left(\mathbf{a}\right) \nabla f_m\left(\mathbf{a}\right)\right. \\ &\quad \left. + f_1\left(\mathbf{a}\right) \nabla g_1\left(\mathbf{a}\right) + \cdots + f_m\left(\mathbf{a}\right) \nabla g_m\left(\mathbf{a}\right)\right]. \end{aligned} \tag{4.102}$$

   The two formulations (4.101) and (4.102) are the same, thus confirming (d).

6. Let $\mathbf{1}_m = (1, 1, \ldots, 1)'$ of length m, so that $\mathbf{1}_m'\mathbf{f}\left(\mathbf{a}\right) = \sum_{i=1}^{m} f_i\left(\mathbf{a}\right)$. Then $\mathbf{D}\left(\mathbf{1}_m'\mathbf{f}\right)\left(\mathbf{a}\right)$ is given by part (d) by taking $\mathbf{g} = (g_1, \ldots, g_m) : A \subset \mathbb{R}^n \to \mathbb{R}^m$ such that each component g_i is identical to one, in which case $\mathbf{D}\mathbf{g}(\mathbf{a}) = \mathbf{0}_{m \times n}$, $\mathbf{g}(\mathbf{a})' = (1, \ldots, 1)$, and $\mathbf{D}\left(\mathbf{1}_m'\mathbf{f}\right)\left(\mathbf{a}\right) = \mathbf{g}(\mathbf{a})'\mathbf{D}\mathbf{f}(\mathbf{a}) = \sum_{i=1}^{m} (\operatorname{grad} f_i)\left(\mathbf{a}\right)$. Observe this result follows from (4.50) and (4.51).

   A bit more generally, let $\mathbf{g} : A \subset \mathbb{R}^n \to \mathbb{R}^m$ be such that each component is identical, given by $\varphi(\mathbf{a}) = g_1(\mathbf{a}) = \cdots = g_m(\mathbf{a})$, where, as above, $\varphi : A \subset \mathbb{R}^n \to \mathbb{R}$. Then its $m \times n$ Jacobian matrix $\mathbf{D}\mathbf{g}(\mathbf{a})$ has identical rows, each being $\nabla\varphi\left(\mathbf{a}\right)$. From (4.102),

$$\begin{aligned}\mathbf{D}\left(\mathbf{f} \cdot \mathbf{g}\right)(\mathbf{a}) &= \mathbf{g}(\mathbf{a})'\mathbf{D}\mathbf{f}(\mathbf{a}) + \mathbf{f}(\mathbf{a})'\mathbf{D}\mathbf{g}(\mathbf{a}) \\ &= \varphi(\mathbf{a}) \sum_{i=1}^{m} (\operatorname{grad} f_i)\left(\mathbf{a}\right) + \nabla\varphi\left(\mathbf{a}\right) \sum_{i=1}^{m} f_i(\mathbf{a}). \end{aligned}$$

Before stating and proving the chain rule on the next page, it is useful to review the definition of differentiability, as given either in (4.89); or directly above in (4.98) and (4.99).

Theorem (The Chain Rule): Let A be an open ball in $\mathbb{R}^n$, and let $\mathbf{f} : A \to \mathbb{R}^m$. Further, let B be an open set in $\mathbb{R}^m$ that contains the range of $\mathbf{f}$, and let $\mathbf{g} : B \to \mathbb{R}^p$. If $\mathbf{f}$ is differentiable at $\mathbf{a} \in A$, and if $\mathbf{g}$ is differentiable at $\mathbf{f}(\mathbf{a})$, then the composition $\mathbf{g} \circ \mathbf{f}$ is differentiable at $\mathbf{a}$, and (notice that the right side represents a product of matrices)

$$\mathbf{D}(\mathbf{g} \circ \mathbf{f})(\mathbf{a}) = \mathbf{D}\mathbf{g}(\mathbf{f}(\mathbf{a}))\mathbf{D}\mathbf{f}(\mathbf{a}). \tag{4.103}$$

Proof: From the differentiability of $\mathbf{f}$ at $\mathbf{a}$ and $\mathbf{g}$ at $\mathbf{f}(\mathbf{a})$, and with $\mathbf{b} = \mathbf{f}(\mathbf{a})$,

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{a}) + \mathbf{D}\mathbf{f}(\mathbf{a})(\mathbf{x} - \mathbf{a}) + \mathbf{r}_\mathbf{f}, \qquad \lim_{\mathbf{x}\to\mathbf{a}} \frac{\mathbf{r}_\mathbf{f}}{\|\mathbf{x} - \mathbf{a}\|} = \mathbf{0}, \tag{4.104}$$

$$\mathbf{g}(\mathbf{y}) = \mathbf{g}(\mathbf{b}) + \mathbf{D}\mathbf{g}(\mathbf{b})(\mathbf{y} - \mathbf{b}) + \mathbf{r}_\mathbf{g}, \qquad \lim_{\mathbf{y}\to\mathbf{b}} \frac{\mathbf{r}_\mathbf{g}}{\|\mathbf{y} - \mathbf{b}\|} = \mathbf{0}, \tag{4.105}$$

It follows from (4.104) that $\mathbf{g}(\mathbf{f}(\mathbf{x})) = \mathbf{g}\big(\mathbf{f}(\mathbf{a}) + \mathbf{D}\mathbf{f}(\mathbf{a})(\mathbf{x} - \mathbf{a}) + \mathbf{r}_\mathbf{f}\big)$, so, with $\mathbf{y} = \mathbf{f}(\mathbf{a}) + \mathbf{D}\mathbf{f}(\mathbf{a})(\mathbf{x} - \mathbf{a}) + \mathbf{r}_\mathbf{f}$, (4.105) implies that

$$\begin{aligned}\mathbf{g}(\mathbf{f}(\mathbf{x})) &= \mathbf{g}(\mathbf{f}(\mathbf{a})) + \mathbf{D}\mathbf{g}(\mathbf{f}(\mathbf{a}))\left(\mathbf{D}\mathbf{f}(\mathbf{a})(\mathbf{x} - \mathbf{a}) + \mathbf{r}_\mathbf{f}\right) + \mathbf{r}_\mathbf{g} \\ &= \mathbf{g}(\mathbf{f}(\mathbf{a})) + \mathbf{D}\mathbf{g}(\mathbf{f}(\mathbf{a}))\mathbf{D}\mathbf{f}(\mathbf{a})(\mathbf{x} - \mathbf{a}) + \mathbf{D}\mathbf{g}(\mathbf{f}(\mathbf{a}))\mathbf{r}_\mathbf{f} + \mathbf{r}_\mathbf{g}.\end{aligned}$$

Thus, it remains to show that

$$\lim_{\mathbf{x}\to\mathbf{a}} \frac{\mathbf{D}\mathbf{g}(\mathbf{f}(\mathbf{a}))\mathbf{r}_\mathbf{f} + \mathbf{r}_\mathbf{g}}{\|\mathbf{x} - \mathbf{a}\|} = \mathbf{0}. \tag{4.106}$$

By definition of $\mathbf{r}_\mathbf{f}$, the first term $\lim_{\mathbf{x}\to\mathbf{a}} \mathbf{D}\mathbf{g}(\mathbf{f}(\mathbf{a}))\mathbf{r}_\mathbf{f} \,/\, \|\mathbf{x} - \mathbf{a}\| = \mathbf{0}$. Further,

$$\lim_{\mathbf{x}\to\mathbf{a}} \mathbf{y} = \lim_{\mathbf{x}\to\mathbf{a}} \big(\mathbf{f}(\mathbf{a}) + \mathbf{D}\mathbf{f}(\mathbf{a})(\mathbf{x} - \mathbf{a}) + \mathbf{r}_\mathbf{f}\big) = \mathbf{f}(\mathbf{a}) = \mathbf{b}.$$

Therefore, $\lim_{\mathbf{x}\to\mathbf{a}} \mathbf{r}_\mathbf{g}/\|\mathbf{y} - \mathbf{b}\| = \mathbf{0}$, i.e.,

$$\lim_{\mathbf{x}\to\mathbf{a}} \frac{\mathbf{r}_\mathbf{g}}{\|\mathbf{D}\mathbf{f}(\mathbf{a})(\mathbf{x} - \mathbf{a}) + \mathbf{r}_\mathbf{f}\|} = \mathbf{0}.$$

Now,

$$\lim_{\mathbf{x}\to\mathbf{a}} \frac{\mathbf{r}_\mathbf{g}}{\|\mathbf{x} - \mathbf{a}\|} = \lim_{\mathbf{x}\to\mathbf{a}} \frac{\mathbf{r}_\mathbf{g}}{\|\mathbf{D}\mathbf{f}(\mathbf{a})(\mathbf{x} - \mathbf{a}) + \mathbf{r}_\mathbf{f}\|} \frac{\|\mathbf{D}\mathbf{f}(\mathbf{a})(\mathbf{x} - \mathbf{a}) + \mathbf{r}_\mathbf{f}\|}{\|\mathbf{x} - \mathbf{a}\|} = \mathbf{0}, \tag{4.107}$$

because (with only the final inequality stated, without justification, in Petrovic) from the triangle inequality and the Generalized Cauchy-Schwarz Inequality (4.135),

$$\begin{aligned}\frac{\|\mathbf{D}\mathbf{f}(\mathbf{a})(\mathbf{x} - \mathbf{a}) + \mathbf{r}_\mathbf{f}\|}{\|\mathbf{x} - \mathbf{a}\|} &\le \frac{\|\mathbf{D}\mathbf{f}(\mathbf{a})(\mathbf{x} - \mathbf{a})\|}{\|\mathbf{x} - \mathbf{a}\|} + \frac{\|\mathbf{r}_\mathbf{f}\|}{\|\mathbf{x} - \mathbf{a}\|} \\ &\le \|\mathbf{D}\mathbf{f}(\mathbf{a})\| + \frac{\|\mathbf{r}_\mathbf{f}\|}{\|\mathbf{x} - \mathbf{a}\|},\end{aligned}$$

the second fraction in (4.107) is bounded, thus showing (4.106) and ending the proof.

4.6.3 The Mean Value Theorem (MVT)

Recall one of the highlights of §4.5 was the MVT (4.81). We now give another proof of the MVT for functions $f : U \subseteq \mathbb{R}^n \to \mathbb{R}$. The proof is very short and easy, because it uses the chain rule, given in both (4.96) and (4.103).

We first review a special case of the chain rule that we will require. In fact, we have it already, in (4.97), but I prefer to redo things (different notation; same concept; good practice). Let $m = 1$ and $p = n$. The chain rule then reads: Let $f : A \subset \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{g} : B \subset \mathbb{R} \to \mathbb{R}^n$ with A and B open sets and $f(A) \subset B$. If f is differentiable at $\mathbf{x} \in A$ and $\mathbf{g}$ is differentiable at $y = f(\mathbf{x})$, then the composite function $(\mathbf{g} \circ f) : A \subset \mathbb{R}^n \to \mathbb{R}^n$ is differentiable at $\mathbf{x} \in A \subset \mathbb{R}^n$, with derivative the $n \times n$ Jacobian matrix

$$\mathbf{J}_{\mathbf{g}\circ f}(\mathbf{x}) = \mathbf{J}_{\mathbf{g}}(f(\mathbf{x}))\,\mathbf{J}_f(\mathbf{x}). \tag{4.108}$$

Given the dimensions of f and $\mathbf{g}$, we can also state the Jacobian of $(f \circ \mathbf{g}) : B \subset \mathbb{R} \to \mathbb{R}$, and now with $\mathbf{y} = \mathbf{g}(x)$ and $x \in B \subset \mathbb{R}$, it is the 1×1 matrix, i.e., scalar

$$J_{f\circ\mathbf{g}}(x) = \mathbf{J}_f(\mathbf{g}(x))\mathbf{J}_{\mathbf{g}}(x), \tag{4.109}$$

this latter case being the one of relevance for the MVT, and the same (albeit with a change in notation) as (4.97). To spell things out for, e.g., $n = 3$, f differentiable at $\mathbf{y} = (y_1, y_2, y_3) \in A$, $\mathbf{g}$ in terms of its component functions from notation (4.8) as $\mathbf{g} = (g_1, g_2, g_3)$, and $\mathbf{g}$ differentiable at $x \in B$, we have

$$\nabla f(\mathbf{y}) = \mathbf{J}_f(\mathbf{y}) = \begin{bmatrix} \frac{\partial f(\mathbf{y})}{\partial y_1} & \frac{\partial f(\mathbf{y})}{\partial y_2} & \frac{\partial f(\mathbf{y})}{\partial y_3} \end{bmatrix}, \quad \mathbf{J}_{\mathbf{g}}(x) = \begin{bmatrix} \frac{\partial g_1(x)}{\partial x} \\ \frac{\partial g_2(x)}{\partial x} \\ \frac{\partial g_3(x)}{\partial x} \end{bmatrix},$$

and

$$\mathbf{J}_{f\circ\mathbf{g}}(x) = \mathbf{J}_f(\mathbf{y})\,\mathbf{J}_{\mathbf{g}}(x) = \sum_{i=1}^{3} \frac{\partial f(\mathbf{y})}{\partial y_i}\frac{\partial g_i(x)}{\partial x}. \tag{4.110}$$

From Terrell, A Passage to Modern Analysis, 2019, p. 316, we have:

Theorem (Mean Value Theorem for Real Functions): Let $f : U \subseteq \mathbb{R}^n \to \mathbb{R}$ and suppose that $\mathbf{a}$ and $\mathbf{b}$ are interior points of U. If the line segment $l_{\mathbf{ab}}$ is contained in the interior of U and f is continuous on $l_{\mathbf{ab}}$ and differentiable at all points of $l_{\mathbf{ab}}$ (except possibly at its endpoints $\mathbf{a}$ and $\mathbf{b}$), then there is a point $\mathbf{c} \in l_{\mathbf{ab}}$ such that

$$f(\mathbf{b}) - f(\mathbf{a}) = \nabla f(\mathbf{c}) \cdot (\mathbf{b} - \mathbf{a}). \tag{4.111}$$

> Proof: The curve $\mathbf{r} : [0,1] \to \mathbb{R}^n$ given by $\mathbf{r}(t) = \mathbf{a} + t(\mathbf{b} - \mathbf{a})$ is continuous on $[0,1]$ and differentiable on $(0,1)$, and $\mathbf{r}'(t) = \mathbf{b} - \mathbf{a}$. Define the function $\phi : [0,1] \to \mathbb{R}^n$ by $\phi(t) = f(\mathbf{r}(t)) = f(\mathbf{a} + t(\mathbf{b} - \mathbf{a}))$. Then ϕ is continuous on $[0,1]$ and differentiable on $(0,1)$, and $\phi(0) = f(\mathbf{a})$, $\phi(1) = f(\mathbf{b})$. Since f is differentiable at all points of $l_{\mathbf{ab}}$, the chain rule applies, and we have
>
> $$\phi'(t) = \nabla f(\mathbf{a} + t(\mathbf{b} - \mathbf{a})) \cdot (\mathbf{b} - \mathbf{a}).$$
>
> By the single variable mean value theorem, there is a $t_0 \in (0,1)$ such that $\phi(1) - \phi(0) = \phi'(t_0)(1 - 0) = \phi'(t_0)$, hence (4.111) holds with $\mathbf{c} = \mathbf{a} + t_0(\mathbf{b} - \mathbf{a})$.

4.7 Higher Order Derivatives and Taylor Series

Recall the notation and results in §4.4. We repeat the basic definitions here, and state some extensions.

Let $f : A \to \mathbb{R}$ with $A \subset \mathbb{R}^n$ an open set such that the partial derivatives $D_i f(\mathbf{x})$ are continuous at point $\mathbf{x} = (x_1, \ldots, x_n) \in A$, $i = 1, 2, \ldots, n$. As $D_i f$ is a function, its partial derivative may be computed, if it exists, i.e., we can apply the D_j operator to $D_i f$ to get $D_j D_i f$, called the iterated partial derivative of f with respect to i and j. As we have shown in §4.4,

If $D_i f$, $D_j f$, $D_i D_j f$ and $D_j D_i f$ exist and are continuous, then

$$D_i D_j f = D_j D_i f. \tag{4.112}$$

This extends as follows. Let $D_i^k f$ denote k applications of D_i to f, e.g., $D_i^2 f = D_i D_i f$. Then, for nonnegative integers k_i, $i = 1, \ldots, n$, any iterated partial derivative operator can be written as $D_1^{k_1} \cdots D_n^{k_n}$, with $D_i^0 = 1$, i.e., the derivative with respect to the ith variable is not taken. The *order* of $D_1^{k_1} \cdots D_n^{k_n}$ is $\sum k_i$. Extending (4.112), if $D_i^j f$ is continuous for $i = 1, \ldots, n$ and $j = 0, \ldots, k_i$, then $D_1^{k_1} \cdots D_n^{k_n} f$ is the same for all possible orderings of the elements of $D_1^{k_1} \cdots D_n^{k_n}$.

Let $f : I \to \mathbb{R}$, with $I \subset \mathbb{R}$ an open interval that contains points x and $x + c$. From (2.283), the Taylor series expansion of $f(c + x)$ around c can be expressed as

$$f(c + x) = \sum_{k=0}^{r} \frac{f^{(k)}(c)}{k!} x^k + \text{ remainder term}, \tag{4.113}$$

if $f^{(r+1)}$ exists, where the order of the expansion is r (before we used n, as is standard convention, but now n is the dimension of the domain of f, which is also standard.) This can be extended to function $f : \mathbb{R}^n \to \mathbb{R}$. We first consider the case with $n = 2$ and $r = 2$, which suffices for many useful applications.

Let $f : A \to \mathbb{R}$ with $A \subset \mathbb{R}^2$ an open set, and let $\mathbf{x} = (x_1, x_2)$ and $\mathbf{c} = (c_1, c_2)$ be column vectors such that $(\mathbf{c} + t\mathbf{x}) \in A$ for $0 \le t \le 1$. Let $g : I \to \mathbb{R}$ be the univariate function defined by $g(t) = f(\mathbf{c} + t\mathbf{x})$ for $I = [0, 1]$, so that $g(0) = f(\mathbf{c})$ and $g(1) = f(\mathbf{c} + \mathbf{x})$. Applying (4.113) to g (with $c = 0$ and $x = 1$) gives

$$g(1) = g(0) + g'(0) + \frac{g''(0)}{2} + + \frac{g'''(0)}{6} + \cdots + \frac{g^{(k)}(0)}{k!} + \cdots . \tag{4.114}$$

From the chain rule (4.97),

$$g'(t) = (\operatorname{grad} f)(\mathbf{c} + t\mathbf{x})\, \mathbf{x} = x_1 D_1 f(\mathbf{c} + t\mathbf{x}) + x_2 D_2 f(\mathbf{c} + t\mathbf{x}) =: f_1(\mathbf{c} + t\mathbf{x}), \tag{4.115}$$

where f_1 is the linear combination of differential operators applied to f given by $f_1 := (x_1 D_1 + x_2 D_2) f$. Again from the chain rule, now applied to f_1, and using (4.112),

$$\begin{aligned} g''(t) &= (\operatorname{grad} f_1)(\mathbf{c} + t\mathbf{x})\, \mathbf{x} \\ &= x_1 D_1 f_1(\mathbf{c} + t\mathbf{x}) + x_2 D_2 f_1(\mathbf{c} + t\mathbf{x}) \\ &= x_1^2 (D_1^2 f)(\mathbf{c} + t\mathbf{x}) + 2 x_1 x_2 (D_1 D_2 f)(\mathbf{c} + t\mathbf{x}) + x_2^2 (D_2^2 f)(\mathbf{c} + t\mathbf{x}) \\ &= \left[(x_1 D_1 + x_2 D_2)^2 f \right](\mathbf{c} + t\mathbf{x}) =: f_2(\mathbf{c} + t\mathbf{x}). \end{aligned}$$

This and (4.115) give $g'(0) = f_1(\mathbf{c})$ and $g''(0) = f_2(\mathbf{c})$, so that (4.114) yields

$$f(\mathbf{c}+\mathbf{x}) = f(\mathbf{c}) + f_1(\mathbf{c}) + \frac{1}{2}f_2(\mathbf{c}) + \cdots \tag{4.116}$$

$$\begin{aligned} &= f(\mathbf{c}) + x_1 D_1 f(\mathbf{c}) + x_2 D_2 f(\mathbf{c}) \\ &\quad + \frac{x_1^2}{2}(D_1^2 f)(\mathbf{c}) + x_1 x_2 (D_1 D_2 f)(\mathbf{c}) + \frac{x_2^2}{2}(D_2^2 f)(\mathbf{c}) + \cdots . \end{aligned} \tag{4.117}$$

If we "remove" the second coordinate used in f, writing x instead of $\mathbf{x} = (x_1, x_2)$ and similar for $\mathbf{c}$, then (4.117) simplifies to

$$f(c+x) = f(c) + xD_1 f(c) + \frac{x^2}{2}(D_1^2 f)(c) + \cdots ,$$

which agrees with (4.113). From (4.114), expansion (4.116) can be continued with

$$\begin{aligned} g'''(t) &= \frac{d}{dt} f_2(\mathbf{c}+t\mathbf{x}) = (\mathrm{grad} f_2)(\mathbf{c}+t\mathbf{x})\,\mathbf{x} = x_1 D_1 f_2(\mathbf{c}+t\mathbf{x}) + x_2 D_2 f_2(\mathbf{c}+t\mathbf{x}) \\ &= x_1 D_1\big(x_1^2(D_1^2 f)(\mathbf{c}+t\mathbf{x}) + 2x_1x_2(D_1D_2 f)(\mathbf{c}+t\mathbf{x}) + x_2^2(D_2^2 f)(\mathbf{c}+t\mathbf{x})\big) \\ &\quad + x_2 D_2\big(x_1^2(D_1^2 f)(\mathbf{c}+t\mathbf{x}) + 2x_1x_2(D_1D_2 f)(\mathbf{c}+t\mathbf{x}) + x_2^2(D_2^2 f)(\mathbf{c}+t\mathbf{x})\big) \\ &= x_1^3(D_1^3 f)(\mathbf{c}+t\mathbf{x}) + 3x_1^2x_2(D_1^2D_2 f)(\mathbf{c}+t\mathbf{x}) \\ &\quad + 3x_1x_2^2(D_1D_2^2 f)(\mathbf{c}+t\mathbf{x}) + x_2^3(D_2^3 f)(\mathbf{c}+t\mathbf{x}) \\ &= \left[(x_1D_1 + x_2D_2)^3 f\right](\mathbf{c}+t\mathbf{x}) \\ &=: f_3(\mathbf{c}+t\mathbf{x}), \end{aligned}$$

and it seems natural to postulate that (4.116) takes the form

$$f(\mathbf{c}+\mathbf{x}) = \sum_{k=0}^{\infty} \frac{f_k(\mathbf{c})}{k!}, \tag{4.118}$$

where $f_k := (x_1D_1 + x_2D_2)^k f$, $k = 0, 1, \ldots$. This is true if all the derivatives exist, and can be proven by induction. Note that f_k can be expanded by the binomial theorem (1.21).

Expression (4.118) is for $n = 2$, although the extension to the case of general n is the same, except that

$$f_k := (x_1D_1 + \cdots + x_nD_n)^k f = (\nabla\mathbf{x})^k f,$$

where we use ∇ to represent the operator that, when applied to f, returns the gradient, i.e., $\nabla := (D_1, \ldots, D_n)$, which is a row vector, and $\mathbf{x} = (x_1, \ldots, x_n)$ is a column vector. Note that $f_1(\mathbf{c})$ is the total differential. Now, f_k is evaluated via the multinomial theorem (see the last part of §1.2), and each f_k will have $\binom{k+n-1}{k}$ terms. With this notation, and assuming all relevant partial derivatives exist, for $f : A \to \mathbb{R}$ with $A \subset \mathbb{R}^n$ an open set,

$$f(\mathbf{c}+\mathbf{x}) = \sum_{k=0}^{r} \frac{\left[(\nabla\mathbf{x})^k f\right](\mathbf{c})}{k!} + \text{ remainder term},$$

where it can be shown (see, e.g., Lang, 1997, §15.5) that the

$$\text{remainder term } = \frac{\left[(\nabla\mathbf{x})^{r+1} f\right](\mathbf{c}+t\mathbf{x})}{(r+1)!}, \quad \text{for} \quad 0 \le t \le 1.$$

4.8 Local Approximation of Real-Valued Multivariate Functions

This section repeats some of the previous material—never a bad thing, unless the goal is absolute efficiency and terseness. It is based on (or, better, a near copy of) Ch. 14 of Fitzpatrick's Advanced Calculus, 2nd ed., 2009, this being a book I discovered after having initially wrote these notes over 20 years ago, and having used the excellent presentations in the books mentioned in the footnote at the beginning of §4.6. Fitzpatrick's presentation is admirable in its detail and clarity, and also covers a few relevant aspects not done in the previous sections. Another, more recent and excellent source of this material, is Petrovic's Advanced Calculus: Theory and Practice, 2nd ed., 2020. Finally, the (in parts magnificent) Terrell, A Passage to Modern Analysis, 2019, offers a slightly more advanced presentation of this material, as well as covering topics not in Fitzpatrick, e.g., Fourier series and Lebesgue integration.

Notes in blue are from me.

Suppose that I is an open interval of real numbers and that the function $f : I \to \mathbb{R}$ is differentiable. By definition, this means that if x is a point in I, then

$$\lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = f'(x).$$

If we rewrite the difference

$$\frac{f(x+h) - f(x)}{h} - f'(x) = \frac{f(x+h) - [f(x) + f'(x)h]}{h},$$

then the above definition of a derivative can be rewritten as

$$\lim_{h\to 0} \frac{f(x+h) - [f(x) + f'(x)h]}{h} = 0. \tag{4.119}$$

Recall (2.285) and (2.286) from the univariate Taylor series expression: Repeating these, if k is a natural number and the function $f : I \to \mathbb{R}$ has continuous derivatives up to order $k+1$, then, for a point x in I and a perturbation $x + h$ that also belongs to I, there is a number θ, $0 < \theta < 1$, such that

$$f(x+h) - \left[f(x) + f'(x)h + \cdots + \left(\frac{1}{k!}\right) f^k(x)h^k\right] = \frac{f^{k+1}(x+\theta h)}{(k+1)!} \cdot h^{k+1}, \tag{4.120}$$

and, therefore, dividing by h^k,

$$\lim_{h\to 0} \frac{f(x+h) - \left[f(x) + f'(x)h + \cdots + (1/k!)f^k(x)h^k\right]}{h^k} = 0. \tag{4.121}$$

In what follows, we wish to establish results analogous to the approximation formulas inherent in (4.119) and, for $k = 2$, in (4.121), for functions of several real variables. It is useful to introduce the following definition.

Definition: Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}$. For a positive integer k, two functions $f : \mathcal{O} \to \mathbb{R}$ and $g : \mathcal{O} \to \mathbb{R}$ are said to be kth-order approximations of one another at the point $\mathbf{x}$, provided that

$$\lim_{\mathbf{h}\to \mathbf{0}} \frac{f(\mathbf{x}+\mathbf{h}) - g(\mathbf{x}+\mathbf{h})}{\|\mathbf{h}\|^k} = 0. \tag{4.122}$$

Example 4.22 *Define $f(h) = e^h$ for each number h. Then $f(0) = f'(0) = f''(0) = 1$. From (4.121) at $x = 0$,*

$$\lim_{h\to 0} \frac{e^h - [1+h]}{h} = 0, \qquad \lim_{h\to 0} \frac{e^h - [1+h+(1/2)h^2]}{h^2} = 0.$$

Thus, the first-degree Taylor polynomial $p_1(h) = 1 + h$ is a first-order approximation of f at $x = 0$, while the second-degree Taylor polynomial $p_2(h) = 1 + h + (1/2)h^2$ is a second-order approximation of f at $x = 0$. ■

The following theorem provides an extension to functions of several variables of the approximation formula (4.119). This is the same as (4.45). Also recall the dot or inner product notation $\langle \cdot, \cdot \rangle$ from (3.1).

Theorem (The First-Order Approximation Theorem): Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ and suppose that the function $f : \mathcal{O} \to \mathbb{R}$ is continuously differentiable. Let $\mathbf{x}$ be a point in $\mathcal{O}$. Then

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{f(\mathbf{x}+\mathbf{h}) - [f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{h}\rangle]}{\|\mathbf{h}\|} = 0. \tag{4.123}$$

Proof: Since $\mathbf{x}$ is an interior point of $\mathcal{O}$, we can choose a positive number r such that the open ball $\mathcal{B}_r(\mathbf{x})$ is contained in $\mathcal{O}$. Fix a nonzero point $\mathbf{h}$ in $\mathbb{R}^n$ with $\|\mathbf{h}\| < r$. Then the point $\mathbf{x}+\mathbf{h}$ belongs to $\mathcal{B}_r(\mathbf{x})$ and so, by the Mean Value Theorem (4.81), we can select a number θ with $0 < \theta < 1$ such that

$$f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) = \langle \nabla f(\mathbf{x}+\theta\mathbf{h}), \mathbf{h}\rangle.$$

Thus,

$$f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{h}\rangle = \langle \nabla f(\mathbf{x}+\theta\mathbf{h}) - \nabla f(\mathbf{x}), \mathbf{h}\rangle,$$

so that, using the Cauchy-Schwarz Inequality, we obtain the estimate

$$|f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{h}\rangle| \le \|\nabla f(\mathbf{x}+\theta\mathbf{h}) - \nabla f(\mathbf{x})\| \cdot \|\mathbf{h}\|.$$

Dividing this estimate by $\|\mathbf{h}\|$, we obtain

$$\frac{|f(\mathbf{x}+\mathbf{h}) - [f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{h}\rangle]|}{\|\mathbf{h}\|} \le \|\nabla f(\mathbf{x}+\theta\mathbf{h}) - \nabla f(\mathbf{x})\|. \tag{4.124}$$

But the function $f : \mathcal{O} \to \mathbb{R}$ has been assumed to be continuously differentiable, so

$$\lim_{\mathbf{h}\to\mathbf{0}} \|\nabla f(\mathbf{x}+\theta\mathbf{h}) - \nabla f(\mathbf{x})\| = 0,$$

and thus (4.123) follows from the estimate (4.124).

For a continuously differentiable function $f : \mathcal{O} \to \mathbb{R}$ whose domain $\mathcal{O}$ is an open subset of the plane $\mathbb{R}^2$ and a point (x_0, y_0) in $\mathcal{O}$, if we denote a general point in $\mathcal{O}$ by (x, y) and set $\mathbf{h} = (x - x_0, y - y_0)$, it is clear that $\mathbf{h}$ approaches $\mathbf{0}$ if and only if (x, y) approaches (x_0, y_0) and that $\|\mathbf{h}\| = \sqrt{(x-x_0)^2 + (y-y_0)^2}$. Hence the approximation property (4.123) can be rewritten as

$$\lim_{(x,y)\to(x_0,y_0)} \frac{f(x,y) - [f(x_0,y_0) + \partial f/\partial x\,(x_0,y_0)\,(x-x_0) + \partial f/\partial y\,(x_0,y_0)\,(y-y_0)]}{\sqrt{(x-x_0)^2 + (y-y_0)^2}} = 0. \tag{4.125}$$

This last formula has a geometric interpretation involving the existence of a tangent plane. To describe this, we state the following definition.

Definition: Let $\mathcal{O}$ be an open subset of the plane $\mathbb{R}^2$ and suppose that the function $f : \mathcal{O} \to \mathbb{R}$ is continuous at the point (x_0, y_0) in $\mathcal{O}$. By the tangent plane to the graph of $f : \mathcal{O} \to \mathbb{R}$ at the point $(x_0, y_0, f(x_0, y_0))$, we mean the graph of a function $\psi : \mathbb{R}^2 \to \mathbb{R}$ of the form

$$\psi(x, y) = a + b(x - x_0) + c(y - y_0) \qquad \text{for } (x, y) \text{ in } \mathbb{R}^2,$$

where a, b, and c are real numbers, which has the property that

$$\lim_{(x,y)\to(x_0,y_0)} \frac{f(x, y) - \psi(x, y)}{\sqrt{(x - x_0)^2 + (y - y_0)^2}} = 0. \tag{4.126}$$

A continuous function of two variables $f : \mathcal{O} \to \mathbb{R}$ can have directional derivatives in all directions at the point (x_0, y_0) in $\mathcal{O}$ without having a tangent plane at the point $(x_0, y_0, f(x_0, y_0))$. Such examples occur because the definition of tangent plane requires that the limit (4.126) exist independently of the way in which the point (x, y) approaches (x_0, y_0). A case in point is Example 4.12.

However, for continuously differentiable functions, the approximation property (4.125) is exactly what is required in order to prove the following corollary. This is the same as (4.45).

Corollary: Suppose that $\mathcal{O}$ is an open subset of the plane $\mathbb{R}^2$ that contains point (x_0, y_0) and that the function $f : \mathcal{O} \to \mathbb{R}$ is continuously differentiable. Then there is a tangent plane to the graph of the function $f : \mathcal{O} \to \mathbb{R}$ at the point $(x_0, y_0, f(x_0, y_0))$. This tangent plane is the graph of the function $\psi : \mathbb{R}^2 \to \mathbb{R}$ defined for (x, y) in $\mathbb{R}^2$ by

$$\psi(x, y) = f(x_0, y_0) + \frac{\partial f}{\partial x}(x_0, y_0)(x - x_0) + \frac{\partial f}{\partial y}(x_0, y_0)(y - y_0). \tag{4.127}$$

Proof: For a general point (x, y) in $\mathcal{O}$, set $\mathbf{h} = (x, y) - (x_0, y_0)$ and observe that

$$\langle \nabla f(x_0, y_0), \mathbf{h} \rangle = \frac{\partial f}{\partial x}(x_0, y_0)(x - x_0) + \frac{\partial f}{\partial y}(x_0, y_0)(y - y_0).$$

Since $f : \mathcal{O} \to \mathbb{R}$ is continuously differentiable, the First-Order Approximation Theorem implies that

$$\lim_{(x,y)\to(x_0,y_0)} \frac{f(x, y) - \psi(x, y)}{\sqrt{(x - x_0)^2 + (y - y_0)^2}} = 0;$$

that is, the graph of the function $\psi : \mathbb{R}^2 \to \mathbb{R}$ is the tangent plane to the graph of the function $f : \mathcal{O} \to \mathbb{R}$ at the point $(x_0, y_0, f(x_0, y_0))$.

We can reason geometrically to see why the tangent plane described in the preceding corollary is necessarily described by equation (4.127). Indeed, suppose that $\mathcal{O}$ is an open subset of the plane $\mathbb{R}^2$ and consider the function $f : \mathcal{O} \to \mathbb{R}$. At the point (x_0, y_0) in $\mathcal{O}$, we look for a plane that is tangent to the graph of $f : \mathcal{O} \to \mathbb{R}$ at the point $(x_0, y_0, f(x_0, y_0))$. If the function $f : \mathcal{O} \to \mathbb{R}$ has first-order partial derivatives at (x_0, y_0), then from the definition of a partial derivative and the meaning, in the case of functions of a single real variable, of the derivative as the slope of the tangent line, it follows that the vectors

$$\mathbf{T}_1 = \left(1, 0, \frac{\partial f}{\partial x}(x_0, y_0)\right) \quad \text{and} \quad \mathbf{T}_2 = \left(0, 1, \frac{\partial f}{\partial y}(x_0, y_0)\right) \tag{4.128}$$

should be parallel to the proposed tangent plane. See Figure 35. Thus, the proposed tangent plane should have a cross-product This is the same as (4.29) and Figure 32.

$$\eta = \mathbf{T}_1 \times \mathbf{T}_2 = \left(-\partial f/\partial x\left(x_0, y_0\right), -\partial f/\partial y\left(x_0, y_0\right), 1\right) \tag{4.129}$$

as a normal vector. The plane that passes through the point $\left(x_0, y_0, f\left(x_0, y_0\right)\right)$ and is normal to η consists of all points (x, y, z) in $\mathbb{R}^3$ that satisfy the equation

$$\langle \eta, (x - x_0, y - y_0, z - f\left(x_0, y_0\right))\rangle = 0,$$

and it is clear that this means that the point (x, y, z) in $\mathbb{R}^3$ lies on the graph of the function defined by equation (4.127).

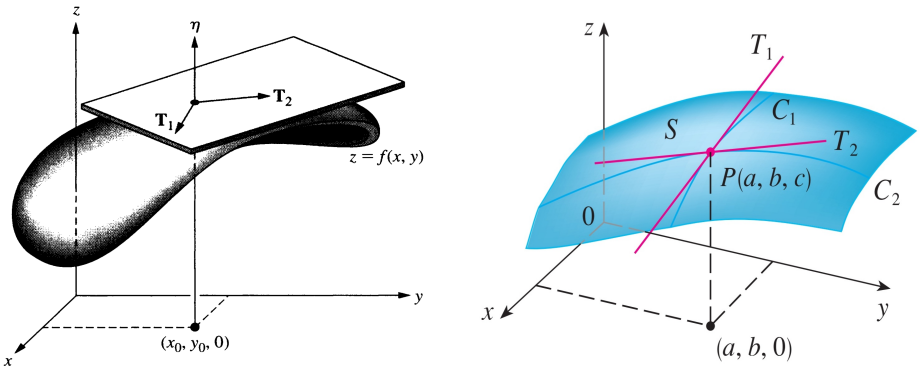


Figure 35: Left: From Fitzpatrick, p. 376: The tangent plane to the graph at the point (x_0, y_0, z_0). Note in (4.128) that $\mathbf{T}_1$, viewed in the xz-plane, can be seen as a vector originating at the origin, with slope $\left(D_1 f\right)\left(x_0, y_0\right)/1$. Similar for $\mathbf{T}_2$ in the yz-plane.
Right: From Stewart, Multivariate Calculus, 7th ed., p. 927: The curve C_1 is the graph of the function $g(x) = f(x, b)$, so the slope of its tangent T_1 at P is $g'(a) = (D_1 f)(a, b)$. The curve C_2 is the graph of the function $h(y) = f(a, y)$, so the slope of its tangent T_2 at P is $h'(b) = (D_2 f)(a, b)$. The partial derivatives $(D_1 f)(a, b)$ and $(D_2 f)(a, b)$ can be interpreted geometrically as the slopes of the tangent lines at $P(a, b, c)$ to the traces C_1 and C_2 of S in the planes $y = b$ and $x = a$.

The First-Order Approximation Theorem is also useful from another, less geometric, perspective. It enables us to approximate rather complicated functions by simpler ones and to assert precisely the manner in which the functions are close to one another. Of course, the simplest type of function is a constant function. The next two simplest types of functions are linear functions and affine functions, which are defined as follows.

Definition: A function $g : \mathbb{R}^n \to \mathbb{R}$ is said to be affine if it is defined by

$$g(\mathbf{u}) = c + \sum_{i=1}^{n} a_i u_i \quad \text{for } \mathbf{u} \text{ in } \mathbb{R}^n,$$

where c and the a_i are prescribed numbers. If $c = 0$, the function is called linear.

Corollary: Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}$ and suppose that the function $f : \mathcal{O} \to \mathbb{R}$ is continuously differentiable. Then there is an affine function that is a first-order approximation of f at the point $\mathbf{x}$, namely, the function $g : \mathbb{R}^n \to \mathbb{R}$ defined by

$$g(\mathbf{u}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{u} - \mathbf{x}\rangle \quad \text{for } \mathbf{u} \text{ in } \mathbb{R}^n.$$

Proof: Observe that the function $g : \mathbb{R}^n \to \mathbb{R}$ is affine and that

$$g(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{h} \rangle \qquad \text{for } \mathbf{x}+\mathbf{h} \text{ in } \mathbb{R}^n.$$

The First-Order Approximation Theorem asserts that the functions $f : \mathcal{O} \to \mathbb{R}$ and $g : \mathcal{O} \to \mathbb{R}$ are first-order approximations of one another at the point $\mathbf{x}$.

Example 4.23 *For $(x, y) \in \mathbb{R}^2$, define $f(x, y) = \sin(x - y - y^2)$. The function $f : \mathbb{R}^2 \to \mathbb{R}$ is continuously differentiable. Computing partial derivatives at the point $(0, 0)$, we find that the affine function that is a first-order approximation of f at the point $(0, 0)$ is defined by $\psi(x, y) = x - y$ for $(x, y) \in \mathbb{R}^2$. Computing partial derivatives at the point $(\pi, 0)$, we find that the affine function that is a first-order approximation of f at the point $(\pi, 0)$ is given by $\psi(x, y) = \pi - x + y$ for $(x, y) \in \mathbb{R}^2$.* ■

Definition: Let $\mathbf{A} = [a_{ij}]$ be an $n \times n$ matrix. The function $Q : \mathbb{R}^n \to \mathbb{R}$ defined by $Q(\mathbf{x}) \equiv \langle \mathbf{Ax}, \mathbf{x} \rangle$, $\mathbf{x} \in \mathbb{R}^n$, is called the quadratic function associated with the matrix $\mathbf{A}$.

We have seen this before, in (3.38). Observe that $\langle \mathbf{Ax}, \mathbf{x} \rangle = (\mathbf{Ax})'\mathbf{x} = \mathbf{xA}'\mathbf{x}$, but $\langle \mathbf{Ax}, \mathbf{x} \rangle = \langle \mathbf{x}, \mathbf{Ax} \rangle = \mathbf{x}'\mathbf{Ax}$, so, without loss of generality, matrix $\mathbf{A}$ can be taken to be symmetric. A non-symmetric matrix $\mathbf{A}$ can be replaced with $\mathbf{B} = (\mathbf{A}' + \mathbf{A})/2$.

Observe that, for $\mathbf{x} \in \mathbb{R}^n$, $Q(\mathbf{x}) = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_j x_i$, so $Q(\mathbf{x})$ is a linear combination of $x_j x_i$'s; hence the name quadratic function.

Definition: Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ and suppose that the function $f : \mathcal{O} \to \mathbb{R}$ has second-order partial derivatives. The Hessian matrix of the function $f : \mathcal{O} \to \mathbb{R}$ at the point $\mathbf{x}$ in $\mathcal{O}$, denoted by $\nabla^2 f(\mathbf{x})$, is defined to be the $n \times n$ matrix that, for each pair of indices i and j, with $1 \leq i \leq n$ and $1 \leq j \leq n$, has the (i, j)th entry defined by

$$\left(\nabla^2 f(\mathbf{x})\right)_{ij} \equiv \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}). \tag{4.130}$$

Observe that, for each index i with $1 \leq i \leq n$,

$$\text{the } i\text{th row of } \nabla^2 f(\mathbf{x}) \text{ is the gradient of the function } \partial f/\partial x_i : \mathcal{O} \to \mathbb{R} \tag{4.131}$$

at the point $\mathbf{x}$.

Also observe that, in view of the equality of cross-partial derivatives, it follows that the Hessian matrix $\nabla^2 f(\mathbf{x})$ is symmetric; that is, the (i, j)th entry equals the (j, i)th entry, provided that the function $f : \mathcal{O} \to \mathbb{R}$ has continuous second-order partial derivatives.

Example 4.24 *Let $\mathcal{O}$ be an open subset of $\mathbb{R}^2$ containing the point (x_0, y_0) and suppose that the function $f : \mathcal{O} \to \mathbb{R}$ has second-order partial derivatives. Then the Hessian matrix of f at (x_0, y_0) is given by*

$$\nabla^2 f(x_0, y_0) = \begin{bmatrix} \partial^2 f/\partial x \partial x\,(x_0, y_0) & \partial^2 f/\partial y \partial x\,(x_0, y_0) \\ \partial^2 f/\partial x \partial y\,(x_0, y_0) & \partial^2 f/\partial y \partial y\,(x_0, y_0) \end{bmatrix}. \qquad ■$$

Theorem: Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}$ and suppose that the function $f : \mathcal{O} \to \mathbb{R}$ has continuous second-order partial derivatives. Choose a positive number r such that the open ball about $\mathbf{x}$, $\mathcal{B}_r(\mathbf{x})$, is contained in $\mathcal{O}$. Then if $\|\mathbf{h}\| < r$ and $|t| \leq 1$, it is useful to recall and review (4.80)

$$\frac{d}{dt}[f(\mathbf{x}+t\mathbf{h})] = \langle \nabla f(\mathbf{x}+t\mathbf{h}), \mathbf{h}\rangle \tag{4.132}$$

and

$$\frac{d^2}{dt^2}[f(\mathbf{x}+t\mathbf{h})] = \left\langle \nabla^2 f(\mathbf{x}+t\mathbf{h})\mathbf{h}, \mathbf{h}\right\rangle. \tag{4.133}$$

Let $\mathbf{H} = \nabla^2 f(\mathbf{x}+t\mathbf{h})$ as in (4.130). Then (4.133) is $\mathbf{h}'\mathbf{H}\mathbf{h}$.

Proof: Let I be an open interval of real numbers that contains the points 0 and 1 and is such that the point $\mathbf{x}+t\mathbf{h}$ belongs to $\mathcal{O}$ if t belongs to I. Define

$$\phi(t) = f(\mathbf{x}+t\mathbf{h}) \quad \text{for } t \text{ in } I.$$

The Directional Derivative Theorem (4.74) implies that, if t is in I, then

$$\phi'(t) = \frac{d}{dt}[f(\mathbf{x}+t\mathbf{h})] = \langle \nabla f(\mathbf{x}+t\mathbf{h}), \mathbf{h}\rangle = \sum_{i=1}^{n} h_i \frac{\partial f}{\partial x_i}(\mathbf{x}+t\mathbf{h}).$$

However, for each index i with $1 \le i \le n$, we can again apply the Directional Derivative Theorem to the partial derivative $\partial f/\partial x_i : \mathcal{O} \to \mathbb{R}$, and, hence, by differentiating each side of the preceding equality, we see that

$$\phi''(t) = \frac{d}{dt}\left[\phi'(t)\right] = \sum_{i=1}^{n} h_i \frac{d}{dt}\left[\frac{\partial f}{\partial x_i}(\mathbf{x}+t\mathbf{h})\right],$$

i.e., and recalling (4.131)

$$\begin{aligned}\phi''(t) &= \sum_{i=1}^{n} h_i \left\langle \nabla\left[\frac{\partial f}{\partial x_i}\right](\mathbf{x}+t\mathbf{h}), \mathbf{h}\right\rangle \\ &= \sum_{i=1}^{n} \left\langle \nabla\left[\frac{\partial f}{\partial x_i}\right](\mathbf{x}+t\mathbf{h}), \mathbf{h}\right\rangle h_i = \left\langle \nabla^2 f(\mathbf{x}+t\mathbf{h})\mathbf{h}, \mathbf{h}\right\rangle.\end{aligned}$$

Definition: The norm of an $n \times n$ matrix $\mathbf{A} = [a_{ij}]$, denoted by $\|\mathbf{A}\|$, is defined by[34]

$$\|\mathbf{A}\| \equiv \sqrt{\sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij}^2}. \tag{4.134}$$

Observe that, if we define the point $\mathbf{A}_i$ in $\mathbb{R}^n$ to be the ith row of the $n \times n$ matrix $\mathbf{A}$, then the square of the norm of $\mathbf{A}$ can be written as

$$\|\mathbf{A}\|^2 = \|\mathbf{A}_1\|^2 + \|\mathbf{A}_2\|^2 + \cdots + \|\mathbf{A}_n\|^2.$$

The above definition of the norm of a matrix is introduced because, with this definition of the norm, we have the following useful variant of the Cauchy-Schwarz Inequality.

[34] The matrix norm defined in (4.134) is one of several popular norms for matrices. An excellent discussion of matrix norms is given by Terrell, A Passage to Modern Analysis, 2019, §9.5.1.

Theorem (A Generalized Cauchy-Schwarz Inequality): Let $\mathbf{A}$ be an $n \times n$ matrix and let $\mathbf{u}$ be a point in $\mathbb{R}^n$. Then

$$\|\mathbf{A}\mathbf{u}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{u}\|. \tag{4.135}$$

Proof: Squaring both sides of (4.135), this inequality holds if and only if

$$\|\mathbf{A}\mathbf{u}\|^2 \leq \|\mathbf{A}\|^2 \|\mathbf{u}\|^2. \tag{4.136}$$

If for each index i with $1 \leq i \leq n$ we let the point $\mathbf{A}_i$ in $\mathbb{R}^n$ be the ith row of $\mathbf{A}$, then

$$\mathbf{A}\mathbf{u} = \left(\langle \mathbf{A}_1, \mathbf{u} \rangle, \ldots, \langle \mathbf{A}_n, \mathbf{u} \rangle\right).$$

Thus, by the standard Cauchy-Schwarz Inequality,

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

We have verified inequality (4.136) and, hence, also inequality (4.135).

Corollary: Let $\mathbf{A}$ be an $n \times n$ matrix, let $Q : \mathbb{R}^n \to \mathbb{R}$ be the quadratic function associated with $\mathbf{A}$, and let $\mathbf{u}$ be a point in $\mathbb{R}^n$. Then

$$|Q(\mathbf{u})| \leq \|\mathbf{A}\| \|\mathbf{u}\|^2. \tag{4.137}$$

Proof: By definition, $|Q(\mathbf{u})| = |\langle \mathbf{A}\mathbf{u}, \mathbf{u} \rangle|$. Thus, if we first use the standard Cauchy-Schwarz Inequality and then the Generalized Cauchy-Schwarz Inequality, it follows that

$$|Q(\mathbf{u})| = |\langle \mathbf{A}\mathbf{u}, \mathbf{u} \rangle| \leq \|\mathbf{A}\mathbf{u}\| \cdot \|\mathbf{u}\| \leq \|\mathbf{A}\| \|\mathbf{u}\|^2.$$

Definition: An $n \times n$ matrix $\mathbf{A}$ is said to be positive definite provided that

$$\langle \mathbf{A}\mathbf{u}, \mathbf{u} \rangle > 0 \quad \text{for all nonzero points } \mathbf{u} \text{ in } \mathbb{R}^n,$$

and is said to be negative definite provided that

$$\langle \mathbf{A}\mathbf{u}, \mathbf{u} \rangle < 0 \quad \text{for all nonzero points } \mathbf{u} \text{ in } \mathbb{R}^n.$$

These are commonly expressed, in shorthand, as $\mathbf{A} > 0$ and $\mathbf{A} < 0$, respectively.

Proposition: Let $\mathbf{A}$ be an $n \times n$ positive definite matrix. Then there is a positive number c such that, for all $\mathbf{u} \in \mathbb{R}^n$,

$$Q(\mathbf{u}) = \langle \mathbf{A}\mathbf{u}, \mathbf{u} \rangle \geq c \|\mathbf{u}\|^2. \tag{4.138}$$

Proof: Since the quadratic function $Q : \mathbb{R}^n \to \mathbb{R}$ is the sum of products of continuous functions, namely, the component projection functions, it is continuous. On the other hand, from (4.4), the unit sphere $S = \{\mathbf{u} \text{ in } \mathbb{R}^n \mid \|\mathbf{u}\| = 1\}$ is a closed and bounded subset of $\mathbb{R}^n$. According to the Sequential Compactness Theorem, the unit sphere is therefore sequentially compact. Thus, by Extreme Value Theorem, there is a point in S that is a minimizer for the restriction of the quadratic function to S. Define c to be the

value of the quadratic function at this minimizer. Observe that c is positive, since we have assumed that the matrix $\mathbf{A}$ is positive definite, and that

$$Q(\mathbf{u}) \geq c \quad \text{for all points } \mathbf{u} \text{ in } S. \tag{4.139}$$

Now, for all points $\mathbf{u}$ in $\mathbb{R}^n$ and all real numbers λ, $\mathbf{A}(\lambda\mathbf{u}) = \lambda\mathbf{A}\mathbf{u}$, so

$$Q(\lambda\mathbf{u}) = \lambda^2 Q(\mathbf{u}). \tag{4.140}$$

Moreover, note that, if $\mathbf{u}$ is any nonzero point in $\mathbb{R}^n$, then

$$Q(\mathbf{u}) = Q\left(\|\mathbf{u}\|\frac{\mathbf{u}}{\|\mathbf{u}\|}\right).$$

From equality (4.140), it follows that

$$Q(\mathbf{u}) = \|\mathbf{u}\|^2 Q\left(\frac{\mathbf{u}}{\|\mathbf{u}\|}\right).$$

But $\mathbf{u}/\|\mathbf{u}\|$ is a point in S, so by inequality (4.139), $Q(\mathbf{u}) \geq c\|\mathbf{u}\|^2$. It is clear that this inequality also holds if $\mathbf{u} = \mathbf{0}$.

Definition: Let $A \subset \mathbb{R}^n$ and $\mathbf{x} \in A$. For function $f : A \to \mathbb{R}$:

i. The point $\mathbf{x}$ is called a local maximizer for the function $f : A \to \mathbb{R}$, provided that there is some positive number r such that

$$f(\mathbf{x}+\mathbf{h}) \leq f(\mathbf{x}) \quad \text{if } \mathbf{x}+\mathbf{h} \text{ is in } A \text{ and } \|\mathbf{h}\| < r.$$

ii. The point $\mathbf{x}$ is called a local minimizer for the function $f : A \to \mathbb{R}$, provided that there is some positive number r such that

$$f(\mathbf{x}+\mathbf{h}) \geq f(\mathbf{x}) \quad \text{if } \mathbf{x}+\mathbf{h} \text{ is in } A \text{ and } \|\mathbf{h}\| < r.$$

iii. The point $\mathbf{x}$ is called a local extreme point for the function $f : A \to \mathbb{R}$, provided that it is either a local minimizer or a local maximizer for $f : A \to \mathbb{R}$.

We immediately find the following necessary condition for a point to be a local extreme point for a function.

Proposition: Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}$ and suppose that the function $f : \mathcal{O} \to \mathbb{R}$ has first-order partial derivatives. If the point $\mathbf{x}$ is a local extreme point for the function $f : \mathcal{O} \to \mathbb{R}$, then

$$\nabla f(\mathbf{x}) = \mathbf{0}. \tag{4.141}$$

Proof: Since $\mathbf{x}$ is an interior point of $\mathcal{O}$, we can choose a positive number r such that the open ball $\mathcal{B}_r(\mathbf{x})$ is contained in $\mathcal{O}$. Fix an index i with $1 \leq i \leq n$ and define the function $\phi : (-r, r) \to \mathbb{R}$ by $\phi(t) = f(\mathbf{x} + t\mathbf{e}_i)$, for $|t| < r$. Then the point 0 is an extreme point of the function $\phi : (-r, r) \to \mathbb{R}$, so recalling (2.68)

$$\phi'(0) = \frac{\partial f}{\partial x_i}(\mathbf{x}) = 0.$$

But this holds for each index i with $1 \leq i \leq n$, which means that (4.141) holds.

Observe that, in order to search for local extreme points, we must first find the points $\mathbf{x}$ in $\mathcal{O}$ at which

$$\nabla f(\mathbf{x}) = \mathbf{0}. \tag{4.142}$$

However, equation (4.142) is a system of n scalar equations in n real unknowns. Unless the function $f : \mathcal{O} \to \mathbb{R}$ has a very simple form, it is not possible to find explicit solutions of (14.18). This should not be so surprising since in fact even for a differentiable function of a single variable $f : \mathbb{R} \to \mathbb{R}$, unless $f : \mathbb{R} \to \mathbb{R}$ is very simple, it is not possible to explicitly find all the numbers x that are solutions of the equation $f'(x) = 0$.

Theorem: Let I be an open interval of real numbers and suppose that the function $f : I \to \mathbb{R}$ has a second derivative. Then for each pair of points x and $x + h$ in the interval I, there is a number θ with $0 < \theta < 1$ such that

$$f(x + h) = f(x) + f'(x)h + \frac{1}{2}f''(x + \theta h)h^2. \tag{4.143}$$

Proof: This is (4.120) for $k = 1$.

From (4.143) for functions of a single variable, and the derivative calculations for functions of several variables we obtained above, we obtain the following theorem.

Theorem: Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ and suppose that the function $f : \mathcal{O} \to \mathbb{R}$ has continuous second-order partial derivatives. For each pair of points $\mathbf{x}$ and $\mathbf{x} + \mathbf{h}$ in $\mathcal{O}$ with the property that the segment between these points also lies in $\mathcal{O}$, there is a number θ with $0 < \theta < 1$ such that

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{h} \rangle + \frac{1}{2} \left\langle \nabla^2 f(\mathbf{x} + \theta\mathbf{h})\mathbf{h}, \mathbf{h} \right\rangle. \tag{4.144}$$

Proof: Choose I to be an open interval of real numbers containing both 0 and 1 such that $\mathbf{x} + t\mathbf{h}$ belongs to $\mathcal{O}$ if t is in I. Then define the function $\psi : I \to \mathbb{R}$ by

$$\psi(t) = f(\mathbf{x} + t\mathbf{h}) \quad \text{for } t \text{ in } I.$$

Recalling (4.133) and (4.135), the function $\psi : I \to \mathbb{R}$ has a second derivative and we have the following formulas for the first and second derivatives, for $t \in I$:

$$\psi'(t) = \langle \nabla f(\mathbf{x} + t\mathbf{h}), \mathbf{h} \rangle \quad \text{and} \quad \psi''(t) = \left\langle \nabla^2 f(\mathbf{x} + t\mathbf{h})\mathbf{h}, \mathbf{h} \right\rangle. \tag{4.145}$$

We now apply (4.143) to the function $\psi : I \to \mathbb{R}$ with $x = 0$ and $h = 1$ to choose a number θ with $0 < \theta < 1$ such that

$$\psi(1) = \psi(0) + \psi'(0) + \frac{1}{2}\psi''(\theta), \tag{4.146}$$

an equality that, after substituting the values of $\psi(1)$ and $\psi(0)$ and using the above formulas for $\psi'(0)$ and $\psi''(\theta)$, is seen to be precisely formula (4.144).

Theorem (The Second-Order Approximation Theorem): Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}$ and suppose that the function $f : \mathcal{O} \to \mathbb{R}$ has continuous second-order partial derivatives. Then

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{f(\mathbf{x}+\mathbf{h}) - [f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{h}\rangle + 1/2 \langle \nabla^2 f(\mathbf{x})\mathbf{h}, \mathbf{h}\rangle]}{\|\mathbf{h}\|^2} = 0. \tag{4.147}$$

Proof: Since the point $\mathbf{x}$ is an interior point of $\mathcal{O}$, we can choose a positive number r such that the open ball $\mathcal{B}_r(\mathbf{x})$ is contained in $\mathcal{O}$. It is convenient to define

$$R(\mathbf{h}) = f(\mathbf{x}+\mathbf{h}) - \left[f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{h}\rangle + \frac{1}{2}\left\langle \nabla^2 f(\mathbf{x})\mathbf{h}, \mathbf{h}\right\rangle\right] \quad \text{for } \|\mathbf{h}\| < r.$$

We must show that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{R(\mathbf{h})}{\|\mathbf{h}\|^2} = 0 \tag{4.148}$$

Fix the point $\mathbf{h}$ in $\mathbb{R}^n$ with $0 < \|\mathbf{h}\| < r$. Using (4.144), we can choose a number θ with $1 < \theta < 1$ such that

$$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{h}\rangle + \frac{1}{2}\left\langle \nabla^2 f(\mathbf{x}+\theta\mathbf{h})\mathbf{h}, \mathbf{h}\right\rangle,$$

so that

$$\begin{aligned} R(\mathbf{h}) &= \frac{1}{2}\left\langle \nabla^2 f(\mathbf{x}+\theta\mathbf{h})\mathbf{h}, \mathbf{h}\right\rangle - \frac{1}{2}\left\langle \nabla^2 f(\mathbf{x}+\mathbf{h})\mathbf{h}, \mathbf{h}\right\rangle \\ &= \frac{1}{2}\left\langle \left[\nabla^2 f(\mathbf{x}+\theta\mathbf{h}) - \nabla^2 f(\mathbf{x}+\mathbf{h})\right]\mathbf{h}, \mathbf{h}\right\rangle. \end{aligned} \tag{4.149}$$

Let $\mathbf{A} = [\nabla^2 f(\mathbf{x}+\theta\mathbf{h}) - \nabla^2 f(\mathbf{x}+\mathbf{h})]$, so that (4.149) is $\mathbf{h}'\mathbf{A}\mathbf{h}$. Then the regular Cauchy-Schwarz implies $|\mathbf{h}'\mathbf{A}\mathbf{h}| \leq \|\mathbf{h}\| \cdot \|\mathbf{A}\mathbf{h}\|$, and the generalized one, (4.135), implies $\|\mathbf{A}\mathbf{h}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{h}\|$. This yields (4.150).

We can use this formula and the Generalized Cauchy-Schwarz Inequality to obtain the estimate

$$|R(\mathbf{h})| \leq \frac{1}{2}\left\|\nabla^2 f(\mathbf{x}+\theta\mathbf{h}) - \nabla^2 f(\mathbf{x})\right\| \|\mathbf{h}\|^2. \tag{4.150}$$

Dividing this estimate by $\|\mathbf{h}\|^2$, we obtain

$$\frac{|R(\mathbf{h})|}{\|\mathbf{h}\|^2} \leq \frac{1}{2}\left\|\nabla^2 f(\mathbf{x}+\theta\mathbf{h}) - \nabla^2 f(\mathbf{x})\right\|. \tag{4.151}$$

But the function $f : \mathcal{O} \to \mathbb{R}$ has been assumed to have continuous second-order partial derivatives, so

$$\lim_{\mathbf{h}\to\mathbf{0}} \left\|\nabla^2 f(\mathbf{x}+\theta\mathbf{h}) - \nabla^2 f(\mathbf{x})\right\| = \mathbf{0},$$

and hence (4.148) follows from the estimate (4.151).

For the following, it is useful to recall the univariate results (2.74) and (2.75).

Theorem (The Second-Derivative Test): Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}$ and suppose that the function $f : \mathcal{O} \to \mathbb{R}$ has continuous second-order partial derivatives. Assume that $\nabla f(\mathbf{x}) = \mathbf{0}$.

i. If the Hessian matrix $\nabla^2 f(\mathbf{x})$ is positive definite, then the point $\mathbf{x}$ is a strict local minimizer of the function $f : \mathcal{O} \to \mathbb{R}$.

ii. If the Hessian matrix $\nabla^2 f(\mathbf{x})$ is negative definite, then the point $\mathbf{x}$ is a strict local maximizer of the function $f : \mathcal{O} \to \mathbb{R}$.

In short,

$$\textbf{If } \nabla^2 f(\mathbf{x}) > 0\textbf{, then } \mathbf{x} \textbf{ is a strict local minimizer of } f. \tag{4.152}$$

$$\textbf{If } \nabla^2 f(\mathbf{x}) < 0\textbf{, then } \mathbf{x} \textbf{ is a strict local maximizer of } f. \tag{4.153}$$

Proof: We need only consider case (i) since case (ii) follows from (i) if we replace f with $-f$. So suppose that the Hessian matrix $\nabla^2 f(\mathbf{x})$ is positive definite. Since the point $\mathbf{x}$ is an interior point of $\mathcal{O}$, we can choose a positive number r such that the open ball $\mathcal{B}_r(\mathbf{x})$ is contained in $\mathcal{O}$. The strategy of the proof is to write the difference $f(\mathbf{x}+\mathbf{h})-f(\mathbf{x})$ as

$$f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) = Q(\mathbf{h}) + R(\mathbf{h}), \tag{4.154}$$

where $Q : \mathbb{R}^n \to \mathbb{R}$ is a positive definite quadratic function and

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{R(\mathbf{h})}{\|\mathbf{h}\|^2} = 0. \tag{4.155}$$

Indeed, if we define for $\|\mathbf{h}\| < r$

$$R(\mathbf{h}) = f(\mathbf{x}+\mathbf{h}) - \left[f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{h} \rangle + \frac{1}{2} \left\langle \nabla^2 f(\mathbf{x})\mathbf{h}, \mathbf{h} \right\rangle \right], \tag{4.156}$$

then the Second-Order Approximation Theorem asserts that (4.155) holds. Moreover, if we define $Q : \mathbb{R}^n \to \mathbb{R}$ to be the quadratic function associated with one-half the Hessian matrix $\nabla^2 f(\mathbf{x})$, then this quadratic function is positive definite. Finally, since $\nabla f(\mathbf{x}) = \mathbf{0}$, we can rewrite (4.156) to obtain (4.154).

Since the quadratic function $Q : \mathbb{R}^n \to \mathbb{R}$ is positive definite, we can use (4.138) to choose a positive number c such that, $\forall \mathbf{h} \in \mathbb{R}^n$, $Q(\mathbf{h}) \geq c\|\mathbf{h}\|^2$. On the other hand, using (4.155), it follows that we can choose a positive number δ less than r such that, for $0 < \|\mathbf{h}\| < \delta$,

$$\frac{|R(\mathbf{h})|}{\|\mathbf{h}\|^2} < \frac{c}{2}, \quad \text{i.e.,} \quad -\frac{c}{2}\|\mathbf{h}\|^2 < R(\mathbf{h}) < \frac{c}{2}\|\mathbf{h}\|^2. \tag{4.157}$$

Combining these two estimates, it follows from (4.154) that, if $0 < \|\mathbf{h}\| < \delta$,

$$f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) = Q(\mathbf{h}) + R(\mathbf{h}) \geq c\|\mathbf{h}\|^2 + R(\mathbf{h}) > \frac{c}{2}\|\mathbf{h}\|^2, \tag{4.158}$$

so the point $\mathbf{x}$ is a strict local minimizer of the function $f : \mathcal{O} \to \mathbb{R}$.

4.9 Approximating Nonlinear Mappings By Linear Mappings

Subsections 4.9.1 and 4.9.2 come from parts of Fitzpatrick, Ch. 15.

4.9.1 Derivative Matrix and Differential

Definition: Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ and consider a mapping $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ represented in component functions as $\mathbf{F} = (F_1, \ldots, F_m)$.

i. The mapping $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is said to have first-order partial derivatives at the point $\mathbf{x}$ in $\mathcal{O}$ provided that for each index i such that $1 \leq i \leq m$, the component function $F_i : \mathcal{O} \to \mathbb{R}$ has first-order partial derivatives at $\mathbf{x}$.

ii. Moreover, the mapping $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is said to have first-order partial derivatives provided that it has first partial derivatives at every point in $\mathcal{O}$.

iii. Finally, the mapping $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is said to be continuously differentiable provided that each component function is continuously differentiable.

Proposition: Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ and suppose that the mapping $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is continuously differentiable. Then the mapping $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is continuous.

> Proof: By definition, each of the component functions of the mapping $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is continuously differentiable. It follows from Theorem (Fitzpatrick, 13.20) (page 229) that each component function is continuous. Consequently, from the Componentwise Continuity Criterion (4.13), we conclude that the mapping $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is itself continuous.

Definition: Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ and suppose that the mapping $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ has first-order partial derivatives at the point $\mathbf{x}$ in $\mathcal{O}$. The **derivative matrix** of $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ at the point $\mathbf{x}$ is defined to be the $m \times n$ matrix $\mathbf{DF}(\mathbf{x})$, which, for each index i such that $1 \leq i \leq m$, has an ith row equal to $\nabla F_i(\mathbf{x})$. Thus, the (i, j)th entry of this derivative matrix is given by the formula

$$(\mathbf{DF}(\mathbf{x}))_{ij} \equiv \frac{\partial F_i}{\partial x_j}(\mathbf{x}). \tag{4.159}$$

Theorem (The Mean Value Theorem for General Mappings): Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ and suppose that the mapping $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is continuously differentiable. Suppose that the points $\mathbf{x}$ and $\mathbf{x} + \mathbf{h}$ are in $\mathcal{O}$ and that the segment joining these points also lies in $\mathcal{O}$. Then there are numbers $\theta_1, \theta_2, \ldots, \theta_m$ in the open interval $(0, 1)$ such that, for $1 \leq i \leq m$,

$$F_i(\mathbf{x} + \mathbf{h}) - F_i(\mathbf{x}) = \langle \nabla F_i \left(\mathbf{x} + \theta_i \mathbf{h}\right), \mathbf{h} \rangle ; \tag{4.160}$$

that is, now necessarily with $\mathbf{x}, \mathbf{h}$ $n \times 1$ column vectors, and $\mathbf{F}(\mathbf{x})$ an $m \times 1$ column vector,

$$\mathbf{F}(\mathbf{x} + \mathbf{h}) - \mathbf{F}(\mathbf{x}) = \mathbf{A}\mathbf{h}, \tag{4.161}$$

where $\mathbf{A}$ is the $m \times n$ matrix whose ith row is $\nabla F_i \left(\mathbf{x} + \theta_i \mathbf{h}\right)$.

> Proof: Just apply the Mean Value Theorem for real-valued functions (4.81) to each of the continuously differentiable component functions and we obtain formula (4.160). Formula (4.161) is simply a rewriting of (4.160).

Recall the First-Order Approximation Theorem for scalar-valued functions, which asserts that, if $\mathcal{O}$ is an open subset of $\mathbb{R}^n$ and the function $f : \mathcal{O} \to \mathbb{R}$ is continuously differentiable, then, at each point $\mathbf{x}$ in $\mathcal{O}$, fixing a typo in Fitzpatrick; and as in (4.123),

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{f(\mathbf{x} + \mathbf{h}) - [f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{h} \rangle]}{\|\mathbf{h}\|} = 0. \tag{4.162}$$

The following is an extension of this result to general mappings.

Theorem (First-Order Approximation Theorem for Mappings): Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}$ and suppose that the mapping $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is continuously differentiable. Then

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\|\mathbf{F}(\mathbf{x}+\mathbf{h}) - [\mathbf{F}(\mathbf{x}) + \mathbf{DF}(\mathbf{x})\mathbf{h}]\|}{\|\mathbf{h}\|} = 0. \tag{4.163}$$

Proof: Since $\mathcal{O}$ is open, we can choose a positive number r such that the open ball $\mathcal{B}_r(\mathbf{x})$ is contained in $\mathcal{O}$. For a point $\mathbf{h}$ in $\mathbb{R}^n$ such that $\|\mathbf{h}\| < r$, define

$$\mathbf{R}(\mathbf{h}) = \mathbf{F}(\mathbf{x}+\mathbf{h}) - [\mathbf{F}(\mathbf{x}) + \mathbf{DF}(\mathbf{x})\mathbf{h}].$$

We must show that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\|\mathbf{R}(\mathbf{h})\|}{\|\mathbf{h}\|} = 0. \tag{4.164}$$

But if we represent the mappings $\mathbf{F}$ and $\mathbf{R}$ as $\mathbf{F} = (F_1, \ldots, F_m)$ and $\mathbf{R} = (R_1, \ldots, R_m)$, then it is clear that, for each index i such that $1 \le i \le m$, and for $\|\mathbf{h}\| < r$,

$$R_i(\mathbf{h}) = F_i(\mathbf{x}+\mathbf{h}) - [F_i(\mathbf{x}) + \langle \nabla F_i(\mathbf{x}), \mathbf{h}\rangle].$$

Since the function $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is continuously differentiable, the First-Order Approximation Theorem for real-valued functions (4.123), also (4.162), implies that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{R_i(\mathbf{h})}{\|\mathbf{h}\|} = 0.$$

Since, for $0 < \|\mathbf{h}\| < r$, taking limits and using result (2.25),

$$\frac{\|\mathbf{R}(\mathbf{h})\|}{\|\mathbf{h}\|} = \left(\sum_{i=1}^{m} \left[\frac{R_i(\mathbf{h})}{\|\mathbf{h}\|} \right]^2 \right)^{1/2},$$

it follows that (4.164) holds.

For a function $f : I \to \mathbb{R}$, where I is an open interval, assume at the point $x \in I$ there is a number a such that Recall (4.119), and also the Lang excerpt starting on page 216.

$$\lim_{h\to 0} \frac{f(x+h) - [f(x) + ah]}{h} = 0.$$

If $h \neq 0$ and $x + h$ is in I,

$$\frac{f(x+h) - [f(x) + ah]}{h} = \frac{f(x+h) - f(x)}{h} - a.$$

It follows that f is differentiable at x and that $f'(x) = a$. This property generalizes to mappings as follows.

Theorem: Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}$ and consider a mapping $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$. Suppose that $\mathbf{A}$ is an $m \times n$ matrix with the property that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\|\mathbf{F}(\mathbf{x}+\mathbf{h}) - [\mathbf{F}(\mathbf{x}) + \mathbf{A}\mathbf{h}]\|}{\|\mathbf{h}\|} = 0. \tag{4.165}$$

Then the mapping $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ has first-order partial derivatives at the point $\mathbf{x}$ and

$$\mathbf{A} = \mathbf{DF}(\mathbf{x}). \tag{4.166}$$

Proof: Represent the mapping $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ in component functions as $\mathbf{F} = (F_1, \ldots, F_m)$ and set $a_{ij} = (\mathbf{A})_{ij}$. We must show that, for each pair of indices i and j such that $1 \leq i \leq m$ and $1 \leq j \leq n$,

$$a_{ij} = \frac{\partial F_i}{\partial x_j}(\mathbf{x}).$$

For each index i such that $1 \leq i \leq m$, define $\mathbf{A}_i$ to be the ith row of the matrix $\mathbf{A}$. Since $\mathcal{O}$ is open, we can choose a positive number r such that the open ball $\mathcal{B}_r(\mathbf{x})$ is contained in $\mathcal{O}$. Now observe that if $1 \leq i \leq m$ and $\|\mathbf{h}\| < r$, then with p_i the ith component projection function (4.5),

$$F_i(\mathbf{x}+\mathbf{h}) - [F_i(\mathbf{x}) + \langle \mathbf{A}_i, \mathbf{h}\rangle] = p_i(\mathbf{F}(\mathbf{x}+\mathbf{h}) - [\mathbf{F}(\mathbf{x}) + \mathbf{A}\mathbf{h}]),$$

so that

$$|F_i(\mathbf{x}+\mathbf{h}) - [F_i(\mathbf{x}) + \langle \mathbf{A}_i, \mathbf{h}\rangle]| \leq \|\mathbf{F}(\mathbf{x}+\mathbf{h}) - [\mathbf{F}(\mathbf{x}) + \mathbf{A}\mathbf{h}]\|.$$

From (4.165) it follows that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{F_i(\mathbf{x}+\mathbf{h}) - [F_i(\mathbf{x}) + \langle \mathbf{A}_i, \mathbf{h}\rangle]}{\|\mathbf{h}\|} = 0.$$

In particular, for an index j such that $1 \leq j \leq n$,

$$\lim_{t\to 0} \frac{F_i(\mathbf{x}+t\mathbf{e}_j) - [F_i(\mathbf{x}) + \langle \mathbf{A}_i, t\mathbf{e}_j\rangle]}{\|t\mathbf{e}_j\|} = 0. \tag{4.167}$$

However, $\|t\mathbf{e}_j\| = |t|$, so (4.167) is equivalent to recall (4.21) or (4.25); and (4.159)

$$\lim_{t\to 0} \frac{F_i(\mathbf{x}+t\mathbf{e}_j) - F_i(\mathbf{x})}{t} = \langle \mathbf{A}_i, \mathbf{e}_j\rangle,$$

thus proving that $\mathbf{F} : \mathcal{O} \to \mathbb{R}^n$ has first partial derivatives at $\mathbf{x}$ and

$$a_{ij} = \langle \mathbf{A}_i, \mathbf{e}_j\rangle = \frac{\partial F_i}{\partial x_j}(\mathbf{x}) \quad \text{for } 1 \leq i \leq m, \;\; 1 \leq j \leq n.$$

The above theorem implies that, for a continuously differentiable mapping, the derivative matrix is the only matrix having the first-order approximation property (4.163).

Definition: Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ that contains the point $\mathbf{x}$ and suppose that the mapping $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ has first-order partial derivatives at the point $\mathbf{x}$. The linear mapping

$$\mathbf{dF}(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^m$$

defined by

$$\mathbf{dF}(\mathbf{x})(\mathbf{h}) \equiv \mathbf{DF}(\mathbf{x})\mathbf{h} \quad \text{for all } \mathbf{h} \text{ in } \mathbb{R}^n$$

is called the **differential of the mapping** $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ at the point $\mathbf{x}$. Compare to the total differential (4.35).

4.9.2 The Chain Rule

The First-Order Approximation Theorem (4.163) is a precise assertion of the manner in which $\mathbf{F}(\mathbf{x}+\mathbf{h}) \approx \mathbf{F}(\mathbf{x}) + \mathbf{dF}(\mathbf{x})(\mathbf{h})$ when $\mathbf{h}$ is sufficiently close to $\mathbf{0}$.

From the Chain Rule for real-valued functions of a single variable, it follows that, if $\mathcal{O}$ and $\mathcal{U}$ are open sets of real numbers and the functions $f : \mathcal{O} \to \mathbb{R}$ and $g : \mathcal{U} \to \mathbb{R}$ are continuously differentiable, with $f(\mathcal{O})$ contained in $\mathcal{U}$, then the composite function $g \circ f : \mathcal{O} \to \mathbb{R}$ is also continuously differentiable and, moreover, for each point x in $\mathcal{O}$, $(g \circ f)'(x) = g'(f(x))f'(x)$.

The Chain Rule carries over to compositions of general continuously differentiable mappings in which the derivative matrix replaces the derivative and matrix multiplication replaces scalar multiplication. The general Chain Rule follows from the following special case of the composition of a mapping with a real-valued function.

In order to clearly state the Chain Rule, it is helpful to use the following notation: For an open subset $\mathcal{U}$ of $\mathbb{R}^m$ and a function $g : \mathcal{U} \to \mathbb{R}$ that has first-order partial derivatives, at each point $\mathbf{p}$ in $\mathcal{U}$, and for each index i such that $1 \leq i \leq m$, we define as in (4.21)

$$D_i g(\mathbf{p}) \equiv \lim_{t \to 0} \frac{g\left(\mathbf{p} + t\mathbf{e}_i\right) - g(\mathbf{p})}{t}. \tag{4.168}$$

This notation has the advantage that the partial derivative with respect to the ith component is denoted by a symbol independent of the notation being used for the points in the domain. Moreover, for each $\mathbf{p} \in \mathcal{U}$, as in (4.22), and is an m-length row vector, i.e., $1 \times m$,

$$\nabla g(\mathbf{p}) = \left(D_1 g(\mathbf{p}), \ldots, D_m g(\mathbf{p})\right). \tag{4.169}$$

Theorem (The Chain Rule): Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ and suppose that the mapping $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is continuously differentiable. Suppose also that $\mathcal{U}$ is an open subset of $\mathbb{R}^m$ and that the function $g : \mathcal{U} \to \mathbb{R}$ is continuously differentiable. Finally, suppose that $\mathbf{F}(\mathcal{O})$ is contained in $\mathcal{U}$. Then the composition $g \circ \mathbf{F} : \mathcal{O} \to \mathbb{R}$ is also continuously differentiable. Moreover, for each point $\mathbf{x}$ in $\mathcal{O}$ and each index i such that $1 \leq i \leq n$,

$$\frac{\partial}{\partial x_i}(g \circ \mathbf{F})(\mathbf{x}) = \sum_{j=1}^{m} D_j g(\mathbf{F}(\mathbf{x})) \frac{\partial F_j}{\partial x_i}(\mathbf{x}); \tag{4.170}$$

that is, with respective matrix sizes $1 \times n$; $1 \times m$; and $m \times n$,

$$\nabla(g \circ \mathbf{F})(\mathbf{x}) = \nabla g(\mathbf{F}(\mathbf{x}))\mathbf{DF}(\mathbf{x}). \tag{4.171}$$

Proof: Let $\mathbf{x}$ be a point in $\mathcal{O}$. Since $\mathcal{O}$ is open, we can select a positive number r such that the open ball $\mathcal{B}_r(\mathbf{x})$ is contained in $\mathcal{O}$. Moreover, since the mapping $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ is continuous and $\mathcal{U}$ is an open subset of $\mathbb{R}^m$, we can also suppose that the segment joining the points $\mathbf{F}(\mathbf{x})$ and $\mathbf{F}(\mathbf{x}+\mathbf{h})$ lies in $\mathcal{U}$ if $\|\mathbf{h}\| < r$. For each $\mathbf{h}$ in $\mathbb{R}^n$ such that $\|\mathbf{h}\| < r$, define

$$\mathbf{R}(\mathbf{h}) = \mathbf{F}(\mathbf{x}+\mathbf{h}) - \mathbf{F}(\mathbf{x}) - \mathbf{DF}(\mathbf{x})\mathbf{h}.$$

According to the First-Order Approximation Theorem for Mappings,

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{\|\mathbf{R}(\mathbf{h})\|}{\|\mathbf{h}\|} = 0, \tag{4.172}$$

and, by the definition of $\mathbf{R}(\mathbf{h})$, if $\|\mathbf{h}\| < r$,

$$\mathbf{F}(\mathbf{x}+\mathbf{h}) - \mathbf{F}(\mathbf{x}) = \mathbf{DF}(\mathbf{x})\mathbf{h} + \mathbf{R}(\mathbf{h}). \tag{4.173}$$

Now for each $\mathbf{h}$ in $\mathbb{R}^n$ such that $\|\mathbf{h}\| < r$, we can apply the MVT (4.81) to the function $g : \mathcal{U} \to \mathbb{R}$ on the segment joining the points $\mathbf{F}(\mathbf{x})$ and $\mathbf{F}(\mathbf{x}+\mathbf{h})$ in order to select a point on this segment, which we label $\mathbf{v}(\mathbf{h})$, at which

$$g(\mathbf{F}(\mathbf{x}+\mathbf{h})) - g(\mathbf{F}(\mathbf{x})) = \langle \nabla g(\mathbf{v}(\mathbf{h})), \mathbf{F}(\mathbf{x}+\mathbf{h}) - \mathbf{F}(\mathbf{x}) \rangle.$$

Substituting (4.173) and using properties (i), (ii), and (iii) of (3.1) gives

$$(g \circ \mathbf{F})(\mathbf{x}+\mathbf{h}) - (g \circ \mathbf{F})(\mathbf{x}) = \langle \nabla g(\mathbf{v}(\mathbf{h})), \mathbf{DF}(\mathbf{x})\mathbf{h} \rangle + \langle \nabla g(\mathbf{v}(\mathbf{h})), \mathbf{R}(\mathbf{h}) \rangle. \tag{4.174}$$

Observe that the continuity of $\mathbf{F} : \mathcal{O} \to \mathbb{R}^m$ implies that

$$\lim_{\mathbf{h}\to\mathbf{0}} \mathbf{v}(\mathbf{h}) = \mathbf{F}(\mathbf{x}). \tag{4.175}$$

We now verify (4.170). Fix an index i, with $1 \le i \le n$. For a number t such that $0 < |t| < r$, if we define $\mathbf{h} = t\mathbf{e}_i$, then from (4.174) we obtain

$$\frac{(g \circ \mathbf{F})(\mathbf{x}+t\mathbf{e}_i) - (g \circ \mathbf{F})(\mathbf{x})}{t} = \langle \nabla g(\mathbf{v}(t\mathbf{e}_i)), \mathbf{DF}(\mathbf{x})\mathbf{e}_i \rangle + \left\langle \nabla g(\mathbf{v}(t\mathbf{e}_i)), \frac{\mathbf{R}(t\mathbf{e}_i)}{t} \right\rangle.$$

From this equality, recalling (4.21), and by using (4.172) and (4.175), it follows that

$$\frac{\partial}{\partial x_i}(g \circ \mathbf{F})(\mathbf{x}) = \langle \nabla g(\mathbf{F}(\mathbf{x})), \mathbf{DF}(\mathbf{x})\mathbf{e}_i \rangle. \tag{4.176}$$

But (note $\mathbf{DF}(\mathbf{x})$ is $m \times n$; $\mathbf{e}_i$ is $n \times 1$, so the rhs is an $m \times 1$ column vector)

$$\mathbf{DF}(\mathbf{x})\mathbf{e}_i = \left(\frac{\partial F_1}{\partial x_i}(\mathbf{x}), \ldots, \frac{\partial F_m}{\partial x_i}(\mathbf{x}) \right),$$

so the scalar equation (4.176) is exactly equation (4.170). In particular, this shows that the function $g \circ \mathbf{F} : \mathcal{O} \to \mathbb{R}$ has first-order partial derivatives, and then, because of the continuity with respect to $\mathbf{x}$ of the right-hand side of formula (4.170), that $g \circ \mathbf{F} : \mathcal{O} \to \mathbb{R}$ is continuously differentiable. To conclude the proof, simply observe that (4.171) is a rewriting of (4.170) in matrix notation.

Example 4.25 *Suppose that the functions $\psi : \mathbb{R}^2 \to \mathbb{R}$ and $\varphi : \mathbb{R}^2 \to \mathbb{R}$ are continuously differentiable. Suppose also that $\mathcal{O}$ is an open subset of the plane $\mathbb{R}^2$ and that the function $f : \mathcal{O} \to \mathbb{R}$ is continuously differentiable. Finally, suppose that $(\psi(x,y), \varphi(x,y))$ is in $\mathcal{O}$ for all (x,y) in $\mathbb{R}^2$. Then*

$$\begin{aligned}\frac{\partial}{\partial x}(f(\psi(x,y),\varphi(x,y))) =& D_1 f(\psi(x,y),\varphi(x,y))\frac{\partial \psi}{\partial x}(x,y) \\ &+ D_2 f(\psi(x,y),\varphi(x,y))\frac{\partial \varphi}{\partial x}(x,y)\end{aligned}$$

and

$$\begin{aligned}\frac{\partial}{\partial y}(f(\psi(x,y),\varphi(x,y))) =& D_1 f(\psi(x,y),\varphi(x,y))\frac{\partial \psi}{\partial y}(x,y) \\ &+ D_2 f(\psi(x,y),\varphi(x,y))\frac{\partial \varphi}{\partial y}(x,y). \qquad \blacksquare\end{aligned}$$

In books in which there are calculations involving partial derivatives, the reader will find a large variety of notation. For example, the second of the two derivative formulas in the previous example is often abbreviated as

$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial \psi}\frac{\partial \psi}{\partial y} + \frac{\partial f}{\partial \varphi}\frac{\partial \varphi}{\partial y}. \tag{4.177}$$

As another common instance of terse but useful notational devices, we note that if the function $f : \mathbb{R}^2 \to \mathbb{R}$ is continuously differentiable and the function $g : \mathbb{R}^2 \to \mathbb{R}$ is defined by

$$g(r, \theta) = f(r\cos\theta, r\sin\theta) \quad \text{for } (r, \theta) \text{ in } \mathbb{R}^2,$$

then according to the Chain Rule, for each point (r, θ) in $\mathbb{R}^2$,

$$\frac{\partial g}{\partial r}(r, \theta) = D_1 f(r\cos\theta, r\sin\theta)\cos\theta + D_2 f(r\cos\theta, r\sin\theta)\sin\theta.$$

The last formula is frequently abbreviated as

$$\frac{\partial f}{\partial r} = \frac{\partial f}{\partial x}\cos\theta + \frac{\partial f}{\partial y}\sin\theta. \tag{4.178}$$

One must carefully interpret this formula in order to understand that it signifies the same thing as its predecessor. Formulas such as (4.177) and (4.178) are useful in compressing long equations. But such formulas are not precise because there is no indication of where the derivatives are to be evaluated, and there is ambiguity about what the variables are. When we analyze functions of two or three variables, especially when computing higher derivatives, it is notationally useful to denote

$$D_1 g(\mathbf{p}) \text{ by } \frac{\partial g}{\partial x}(\mathbf{p}), \quad D_2 g(\mathbf{p}) \text{ by } \frac{\partial g}{\partial y}(\mathbf{p}), \quad \text{and} \quad D_3 g(\mathbf{p}) \text{ by } \frac{\partial g}{\partial z}(\mathbf{p}),$$

even when x, y, and z have not been explicitly introduced as notation for the component variables. In the following example, we use this notational convention.

We make some notes in preparation for the next example: Function $u : \mathbb{R}^2 \to \mathbb{R}$ is not specified, but we are told it is harmonic. We label its inputs as $(x, y) \in \mathbb{R}^2$. Let function $\mathbf{H} : \mathbb{R}^2 \to \mathbb{R}^2$ be given by $\mathbf{H}(w, z) = (H_1, H_2) = (w^2 - z^2, 2wz)$, and define the composite function $v : \mathbb{R}^2 \to \mathbb{R}$ as $v = (u \circ \mathbf{H})$. Then $v(x, y) = (u \circ \mathbf{H})(x, y) = u(x^2 - y^2, 2xy)$ and, with $\mathbf{x} = (x, y)$, (4.171) is

$$\begin{aligned}
\left[\ \tfrac{\partial v}{\partial x}(\mathbf{x}) \quad \tfrac{\partial v}{\partial y}(\mathbf{x})\ \right] &= \nabla(u \circ \mathbf{H})(\mathbf{x}) = \nabla u(\mathbf{H}(\mathbf{x}))\,\mathbf{DH}(\mathbf{x}) \\
&= \left[\ (D_1 u)(\mathbf{H}(\mathbf{x})) \quad (D_2 u)(\mathbf{H}(\mathbf{x}))\ \right] \begin{bmatrix} \frac{\partial H_1}{\partial x}(\mathbf{x}) & \frac{\partial H_1}{\partial y}(\mathbf{x}) \\ \frac{\partial H_2}{\partial x}(\mathbf{x}) & \frac{\partial H_2}{\partial y}(\mathbf{x}) \end{bmatrix} \\
&= \left[\ (D_1 u)(x^2 - y^2, 2xy) \quad (D_2 u)(x^2 - y^2, 2xy)\ \right] \begin{bmatrix} 2x & 2y \\ 2y & 2x \end{bmatrix},
\end{aligned}$$

and, writing out the two rows separately (equivalently as in (4.170)),

$$\begin{aligned}
\frac{\partial v}{\partial x}(\mathbf{x}) &= \frac{\partial}{\partial x}(u \circ \mathbf{H})(\mathbf{x}) = (D_1 u)(\mathbf{H}(\mathbf{x}))\frac{\partial H_1}{\partial x}(\mathbf{x}) + (D_2 u)(\mathbf{H}(\mathbf{x}))\frac{\partial H_2}{\partial x}(\mathbf{x}), \\
\frac{\partial v}{\partial y}(\mathbf{x}) &= \frac{\partial}{\partial y}(u \circ \mathbf{H})(\mathbf{x}) = (D_1 u)(\mathbf{H}(\mathbf{x}))\frac{\partial H_1}{\partial y}(\mathbf{x}) + (D_2 u)(\mathbf{H}(\mathbf{x}))\frac{\partial H_2}{\partial y}(\mathbf{x}).
\end{aligned}$$

Using now the informal notation,

$$\begin{aligned}\frac{\partial v}{\partial x}(x,y) &= \frac{\partial u}{\partial x}\left(x^2-y^2,2xy\right)2x+\frac{\partial u}{\partial y}\left(x^2-y^2,2xy\right)2y,\\ \frac{\partial v}{\partial y}(x,y) &= \frac{\partial u}{\partial x}\left(x^2-y^2,2xy\right)2y+\frac{\partial u}{\partial y}\left(x^2-y^2,2xy\right)2x.\end{aligned}$$

The following example also involves computing second derivatives as in (4.130).

Example 4.26 *(Fitzpatrick, p. 417) A function $u:\mathbb{R}^2\to\mathbb{R}$ is said to be harmonic provided it has continuous second-order partial derivatives that satisfy the identity*

$$\frac{\partial^2 u}{\partial x^2}(x,y)+\frac{\partial^2 u}{\partial y^2}(x,y)=0 \quad \text{for all } (x,y) \text{ in } \mathbb{R}^2.$$

Suppose that the function $u:\mathbb{R}^2\to\mathbb{R}$ is harmonic. Define

$$v(x,y)=u\left(x^2-y^2,2xy\right) \quad \text{for all } (x,y) \text{ in } \mathbb{R}^2.$$

Then it turns out that the function $v:\mathbb{R}^2\to\mathbb{R}$ is also harmonic. To verify this, we must show that

$$\frac{\partial^2 v}{\partial x^2}(x,y)+\frac{\partial^2 v}{\partial y^2}(x,y)=0 \quad \text{for all } (x,y) \text{ in } \mathbb{R}^2.$$

However, for (x,y) in $\mathbb{R}^2$,

$$\frac{\partial v}{\partial x}(x,y)=\frac{\partial u}{\partial x}\left(x^2-y^2,2xy\right)2x+\frac{\partial u}{\partial y}\left(x^2-y^2,2xy\right)2y,$$

So

$$\begin{aligned}\frac{\partial^2 v}{\partial x^2}(x,y) =& \frac{\partial^2 u}{\partial x^2}\left(x^2-y^2,2xy\right)4x^2+\frac{\partial u}{\partial x}\left(x^2-y^2,2xy\right)2\\ &+\frac{\partial^2 u}{\partial x\partial y}\left(x^2-y^2,2xy\right)8xy+\frac{\partial^2 u}{\partial y^2}\left(x^2-y^2,2xy\right)4y^2.\end{aligned}$$

We carry out a similar computation for $\partial^2 v/\partial y^2(x,y)$, and since

$$\frac{\partial^2 u}{\partial x^2}\left(x^2-y^2,2xy\right)+\frac{\partial^2 u}{\partial y^2}\left(x^2-y^2,2xy\right)=0 \quad \text{for all } (x,y) \text{ in } \mathbb{R}^2,$$

a calculation shows that

$$\frac{\partial^2 v}{\partial x^2}(x,y)+\frac{\partial^2 v}{\partial y^2}(x,y)=0 \quad \text{for } (x,y) \text{ in } \mathbb{R}^2. \qquad \blacksquare$$

The special case of the Chain Rule that we have just proved leads to the proof of the general case.

Theorem (The Chain Rule for General Mappings): Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ and suppose that the mapping $\mathbf{F}:\mathcal{O}\to\mathbb{R}^m$ is continuously differentiable. Suppose also that $\mathcal{U}$ is an open subset of $\mathbb{R}^m$ and that the mapping $\mathbf{G}:\mathcal{U}\to\mathbb{R}^k$ is continuously differentiable. Finally, suppose that $\mathbf{F}(\mathcal{O})$ is contained in $\mathcal{U}$. Then the composite mapping $\mathbf{G}\circ\mathbf{F}:\mathcal{O}\to\mathbb{R}^k$ is also continuously differentiable. Moreover, for each point $\mathbf{x}$ in $\mathcal{O}$,
with respective matrix sizes $k\times n$, $k\times m$, and $m\times n$,

$$\mathbf{D}(\mathbf{G}\circ\mathbf{F})(\mathbf{x})=\mathbf{DG}(\mathbf{F}(\mathbf{x}))\cdot\mathbf{DF}(\mathbf{x}). \tag{4.179}$$

Proof: Represent the mapping $\mathbf{G}$ in component functions by $\mathbf{G} = (G_1, \ldots, G_k)$. Then observe that the composition $\mathbf{G} \circ \mathbf{F} : \mathcal{O} \to \mathbb{R}^k$ is represented in component functions by $\mathbf{G} \circ \mathbf{F} = (G_1 \circ \mathbf{F}, G_2 \circ \mathbf{F}, \ldots, G_k \circ \mathbf{F})$. For an index j such that $1 \leq j \leq k$, and that the component function $G_j : \mathcal{U} \to \mathbb{R}$ is continuously differentiable, recall the Componentwise Continuity Criterion (4.13) it follows from the Chain Rule (4.170) and (4.171) that, for all points $\mathbf{x}$ in $\mathcal{O}$,

$$\nabla \left(G_j \circ \mathbf{F} \right) (\mathbf{x}) = \nabla G_j(\mathbf{F}(\mathbf{x})) \mathbf{DF}(\mathbf{x}).$$

This formula is an assertion of the equality of the jth rows of each of the matrices in formula (4.179) for $1 \leq j \leq k$. Thus, the matrix formula (4.179) holds. Therefore, the composition $\mathbf{G} \circ \mathbf{F} : \mathcal{O} \to \mathbb{R}^k$ has first-order partial derivatives at each point, and from the continuity of the entries on the right-hand side of (4.179), again from (4.13) we conclude that the composition is continuously differentiable.

4.9.3 Directional Derivatives

In (4.73), we defined the directional derivative of the function $f : \mathcal{O} \subset \mathbb{R}^n \to \mathbb{R}$ in the direction $\mathbf{p}$ at the point $\mathbf{x}$. Further, in (4.74) and (4.80), we expressed $(D_{\mathbf{p}} f)(\mathbf{x})$ as $(\nabla f)(\mathbf{x}) \cdot \mathbf{p}$, this being a matrix product of $1 \times n$ and $n \times 1$ vectors.

This extends in a natural way to the case of $\mathbf{F} : \mathcal{O} \subset \mathbb{R}^n \to \mathbb{R}^m$ continuously differentiable, with derivative matrix $\mathbf{DF}(\mathbf{x})$ from (4.159). In particular, for $\mathbf{p}, \mathbf{x} \in \mathcal{O}$, $\mathbf{D_p F}(\mathbf{x})$ is the $m \times 1$ vector $\mathbf{DF}(\mathbf{x}) \cdot \mathbf{p}$.

Example 4.27 *(Dineen, Multivariate Calculus and Geometry, 3rd ed., p. 8) Let* $\mathbf{F} : \mathbb{R}^4 \to \mathbb{R}^3$ *be defined by*

$$\mathbf{F}(x, y, z, w) = \left(x^2 y, xyz, x^2 + y^2 + zw^2 \right).$$

Then $\mathbf{F} = (f_1, f_2, f_3)$*, where* $f_1(x, y, z, w) = x^2 y$*,* $f_2(x, y, z, w) = xyz$ *and* $f_3(x, y, z, w) = x^2 + y^2 + zw^2$*. Moreover,* $\nabla f_1(x, y, z, w) = (2xy, x^2, 0, 0)$*,* $\nabla f_2(x, y, z, w) = (yz, xz, xy, 0)$ *and* $\nabla f_3(x, y, z, w) = (2x, 2y, w^2, 2zw)$*. Hence*

$$\mathbf{DF}(x, y, z, w) = \begin{pmatrix} 2xy & x^2 & 0 & 0 \\ yz & xz & xy & 0 \\ 2x & 2y & w^2 & 2zw \end{pmatrix}.$$

If $\mathbf{x} = (1, 2, -1, -2)'$ *and* $\mathbf{p} = (0, 1, 2, -2)'$ *then*

$$\mathbf{D_p F}(\mathbf{x}) = \mathbf{DF}(\mathbf{x}) \cdot \mathbf{p} = \begin{pmatrix} 4 & 1 & 0 & 0 \\ -2 & -1 & 2 & 0 \\ 2 & 4 & 4 & 4 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 2 \\ -2 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 4 \end{pmatrix}. \qquad \blacksquare$$

5 Multivariate Integration

> Mathematics is not a deductive science – that's a cliche. When you try to prove a theorem, you don't just list the hypotheses, and then start to reason. What you do is trial and error, experimentation, guesswork. (Paul R. Halmos)

5.1 Definitions, Existence, and Properties

This subsection is based on Fitzpatrick, §18.1. The quote from Pugh at the beginning of §4.2 pertains precisely to this material, i.e., "The multivariable case in which $f : \mathbb{R}^n \to \mathbb{R}$ offers no new ideas, only new notation", compared to the univariate case in §2.4.1. Still, it is worth spelling out, and also serves as a refresher of the univariate material.

Recall from §2.4.1 that, if $I = [a, b]$, $a < b$, is a closed bounded interval of real numbers, m is a positive integer, and $P = \{x_0, \ldots, x_m\}$ are $m + 1$ real numbers such that

$$a = x_0 < x_1 < \cdots < x_i < \cdots < x_m = b, \tag{5.1}$$

then P is called a partition of $[a, b]$, and the intervals $[x_{i-1}, x_i]$, for i an index between 1 and m, are called intervals in the partition P. We define the length of the interval $I = [a, b]$ to be $b - a$. Let n be a positive integer and for each index i between 1 and n let $I_i = [a_i, b_i]$ be a closed bounded interval of real numbers. The Cartesian product of these intervals,

$$\mathbf{I} = I_1 \times \cdots \times I_i \times \cdots \times I_n = \{\mathbf{x} = (x_1, \ldots, x_n) \text{ in } \mathbb{R}^n \mid x_i \text{ in } I_i \text{ for } 1 \le i \le n\}, \tag{5.2}$$

is called a **generalized rectangle**. It is convenient to refer to the interval I_i as being the ith edge of $\mathbf{I}$. We define the volume of $\mathbf{I}$, denoted by vol $\mathbf{I}$, to be the product of the lengths of the n edges; that is,

$$\operatorname{vol} \mathbf{I} \equiv \prod_{i=1}^{n} [b_i - a_i].$$

In the case where $n = 1$, the volume is simply the length; in the case where $n = 2$, the volume is called the area.

Definition: Given a generalized rectangle $\mathbf{I} = I_1 \times \cdots \times I_i \times \cdots \times I_n$, for each index i between 1 and n, let P_i be a partition of the ith edge I_i. The collection of generalized rectangles of the form

$$\mathbf{J} = J_1 \times \cdots \times J_i \times \cdots \times J_n,$$

where each J_i is an interval in the partition P_i, is called a partition of $\mathbf{I}$ and is denoted by

$$\mathbf{P} \equiv (P_1, \ldots, P_n).$$

Consider the rectangle $[a, b] \times [c, d]$ in the plane $\mathbb{R}^2$. Let $P_1 = \{x_0, \ldots, x_m\}$ and $P_2 = \{y_0, \ldots, y_\ell\}$ be partitions of $[a, b]$ and $[c, d]$, respectively, and define $\mathbf{P} = (P_1, P_2)$. Then

$$\begin{aligned}\sum_{\mathbf{J} \text{ in } \mathbf{P}} \operatorname{vol} \mathbf{J} &= \sum_{j=1}^{\ell} \sum_{i=1}^{m} [x_i - x_{i-1}] [y_j - y_{j-1}] = \sum_{j=1}^{\ell} \left\{ \sum_{i=1}^{m} [x_i - x_{i-1}] \right\} [y_j - y_{j-1}] \\ &= \sum_{j=1}^{\ell} \{[b - a]\} [y_j - y_{j-1}] = [b - a] \sum_{j=1}^{\ell} [y_j - y_{j-1}] = [b - a][d - c] = \operatorname{vol} \mathbf{I}.\end{aligned}$$

An induction argument shows that the above formula also holds in general: For each natural number n, if $\mathbf{P}$ is a partition of the generalized rectangle $\mathbf{I}$ in $\mathbb{R}^n$, then

$$\operatorname{vol} \mathbf{I} = \sum_{\mathbf{J} \text{ in } \mathbf{P}} \operatorname{vol} \mathbf{J}. \tag{5.3}$$

Let $f : \mathbf{I} \to \mathbb{R}$ is a bounded function whose domain $\mathbf{I}$ is a generalized rectangle and let $\mathbf{P}$ be a partition of $\mathbf{I}$. For $\mathbf{J}$ a generalized rectangle in the partition $\mathbf{P}$, we define

$$m(f, \mathbf{J}) \equiv \inf\{f(\mathbf{x}) \mid \mathbf{x} \text{ in } \mathbf{J}\} \quad \text{and} \quad M(f, \mathbf{J}) \equiv \sup\{f(\mathbf{x}) \mid \mathbf{x} \text{ in } \mathbf{J}\}.$$

Remark: Note that $\mathbf{I}$ is closed and bounded. If f is continuous, then, for $n = 1$, from (2.27), it is uniformly continuous. From the EVT (4.19), the inf and sup can be replaced with min and max.

We then define the lower and upper Darboux sums for the function $f : \mathbf{I} \to \mathbb{R}$ with respect to the partition $\mathbf{P}$, by

$$L(f, \mathbf{P}) \equiv \sum_{\mathbf{J} \text{ in } \mathbf{P}} m(f, \mathbf{J}) \text{ vol } \mathbf{J}, \quad \text{and} \quad U(f, \mathbf{P}) \equiv \sum_{\mathbf{J} \text{ in } \mathbf{P}} M(f, \mathbf{J}) \text{ vol } \mathbf{J}.$$

Lemma: Let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function on a generalized rectangle $\mathbf{I}$. Suppose that the two numbers m and M have the property that $\forall \mathbf{x} \in \mathbf{I}$, $m \leq f(\mathbf{x}) \leq M$. Then, for any partition $\mathbf{P}$ of $\mathbf{I}$,

$$m \text{ vol } \mathbf{I} \leq L(f, \mathbf{P}) \leq U(f, \mathbf{P}) \leq M \text{ vol } \mathbf{I}. \tag{5.4}$$

Proof: Let $\mathbf{P}$ be a partition of $\mathbf{I}$. For a generalized rectangle $\mathbf{J}$ in $\mathbf{I}$, it is clear that

$$m \leq \inf\{f(\mathbf{x}) \mid \mathbf{x} \text{ in } \mathbf{J}\} = m(f, \mathbf{J}) \leq M(f, \mathbf{J}) = \sup\{f(\mathbf{x}) \mid \mathbf{x} \text{ in } \mathbf{J}\} \leq M,$$

so

$$m \text{ vol } \mathbf{J} \leq m(f, \mathbf{J}) \text{ vol } \mathbf{J} \leq M(f, \mathbf{J}) \text{ vol } \mathbf{J} \leq M \text{ vol } \mathbf{J}.$$

Summing over all the generalized rectangles $\mathbf{J}$ in the partition $\mathbf{P}$ and using the sum of volumes formula (5.3), we conclude that the inequality (5.4) holds.

Given a partition $\mathbf{P} = (P_1, \ldots, P_n)$ of a generalized rectangle $\mathbf{I}$, another partition $\mathbf{P}^* = (P_1^*, \ldots, P_n^*)$ of $\mathbf{I}$ is said to be a refinement of $\mathbf{P}$ provided that, for each index i between 1 and n, P_i^* is a refinement of P_i. Recall (2.123). Observe that if $\mathbf{P}^*$ is a refinement of $\mathbf{P}$, then (i) each generalized rectangle $\mathbf{J}$ in $\mathbf{P}^*$ is contained in exactly one generalized rectangle in $\mathbf{P}$, and (ii) given a generalized rectangle $\mathbf{J}$ in $\mathbf{P}$, the collection of generalized rectangles in $\mathbf{P}^*$ contained in $\mathbf{J}$ induces a partition of $\mathbf{J}$ that we denote by $\mathbf{P}^*(\mathbf{J})$. The following distribution formulas for the lower and upper Darboux sums follow from these two properties:

$$L(f, \mathbf{P}^*) = \sum_{\mathbf{J} \text{ in } \mathbf{P}} L(f, \mathbf{P}^*(\mathbf{J})) \quad \text{and} \quad U(f, \mathbf{P}^*) = \sum_{\mathbf{J} \text{ in } \mathbf{P}} U(f, \mathbf{P}^*(\mathbf{J})). \tag{5.5}$$

Lemma (The Refinement Lemma): Let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function on a generalized rectangle $\mathbf{I}$. Let $\mathbf{P}$ be a partition of $\mathbf{I}$ and let $\mathbf{P}^*$ be a refinement of $\mathbf{P}$. Then

$$L(f, \mathbf{P}) \leq L(f, \mathbf{P}^*) \leq U(f, \mathbf{P}^*) \leq U(f, \mathbf{P}). \tag{5.6}$$

Proof: Let $\mathbf{J}$ be a generalized rectangle in $\mathbf{P}$ and denote by $\mathbf{P}^*(\mathbf{J})$ the partition of $\mathbf{J}$ induced by $\mathbf{P}^*$. From (5.4), with $\mathbf{J}$ playing the role of $\mathbf{I}$, it follows that

$$m(f, \mathbf{J}) \text{ vol } \mathbf{J} \leq L\left(f, \mathbf{P}^*(\mathbf{J})\right) \leq U\left(f, \mathbf{P}^*(\mathbf{J})\right) \leq M(f, \mathbf{J}) \text{ vol } \mathbf{J}.$$

If we sum these inequalities over all generalized rectangles $\mathbf{J}$ in $\mathbf{P}$ and use the distribution formulas (5.5), we arrive at the inequality (5.6).

For two partitions P and P' of a closed bounded interval of real numbers I, by taking the partition consisting of all points that are partition points in at least one of the two partitions, we obtain a partition that is a common refinement of the two given partitions, meaning that it is a refinement of both P and P'. Similarly, suppose that $\mathbf{P}$ and $\mathbf{P}'$ are two partitions of a generalized rectangle $\mathbf{I}$ in $\mathbb{R}^n$ represented as $\mathbf{P} = (P_1, \ldots, P_n)$ and $\mathbf{P}' = (P_1', \ldots, P_n')$. For each index i between 1 and n, choose P_i'' to be a common refinement of P_i and P_i' and define $\mathbf{P}'' = (P_1'', \ldots, P_n'')$. Then $\mathbf{P}''$ is a partition of $\mathbf{I}$ that is a common refinement of the partitions $\mathbf{P}$ and $\mathbf{P}'$. The existence of common refinements is what is necessary to establish the following proposition.

Proposition: Let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function on a generalized rectangle $\mathbf{I}$. For any two partitions $\mathbf{P}_1$ and $\mathbf{P}_2$ of $\mathbf{I}$,

$$L\left(f, \mathbf{P}_1\right) \leq U\left(f, \mathbf{P}_2\right). \tag{5.7}$$

Proof: Choose $\mathbf{P}$ to be a common refinement of the two partitions $\mathbf{P}_1$ and $\mathbf{P}_2$. By the Refinement Lemma,

$$L\left(f, \mathbf{P}_1\right) \leq L(f, \mathbf{P}) \leq U(f, \mathbf{P}) \leq U\left(f, \mathbf{P}_2\right).$$

Definition: Let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function on a generalized rectangle $\mathbf{I}$. We define the lower and upper integrals of f on $\mathbf{I}$, by

$$\underline{\int_{\mathbf{I}}} f \equiv \sup\{L(f, \mathbf{P}) \mid \mathbf{P} \text{ a partition of the generalized rectangle } \mathbf{I}\} \tag{5.8}$$

and

$$\overline{\int_{\mathbf{I}}} f \equiv \inf\{U(f, \mathbf{P}) \mid \mathbf{P} \text{ a partition of the generalized rectangle } \mathbf{I}\}. \tag{5.9}$$

Lemma: Let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function on a generalized rectangle $\mathbf{I}$. Then

$$\underline{\int_{\mathbf{I}}} f \leq \overline{\int_{\mathbf{I}}} f.$$

Proof: Let $\mathbf{P}$ be a partition of $\mathbf{I}$. Proposition (5.7) asserts that $U(f, \mathbf{P})$ is an upper bound for the collection of all lower Darboux sums for f. Therefore, by the definition of supremum,

$$\underline{\int_{\mathbf{I}}} f \leq U(f, \mathbf{P}).$$

But this inequality asserts that $\int_{\mathbf{I}} f$ is a lower bound for the collection of upper Darboux sums for f. Thus, by the definition of infimum,

$$\underline{\int_{\mathbf{I}}} f \leq \overline{\int_{\mathbf{I}}} f.$$

Definition: Let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function on a generalized rectangle $\mathbf{I}$. Then we say that $f : \mathbf{I} \to \mathbb{R}$ is integrable, or f is integrable on $\mathbf{I}$, provided that

$$\underline{\int_{\mathbf{I}}} f = \overline{\int_{\mathbf{I}}} f. \tag{5.10}$$

When this is so, the integral of the function $f : \mathbf{I} \to \mathbb{R}$, denoted by $\int_{\mathbf{I}} f$, is defined by

$$\int_{\mathbf{I}} f \equiv \underline{\int_{\mathbf{I}}} f = \overline{\int_{\mathbf{I}}} f, \quad \text{and we write} \quad f \in \mathcal{R}[\mathbf{I}]. \tag{5.11}$$

Definition: Let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function on a generalized rectangle. For each natural number k, let $\mathbf{P}_k$ be a partition of $\mathbf{I}$. The sequence of partitions $\{\mathbf{P}_k\}$ is said to be an Archimedean sequence of partitions for the function $f : \mathbf{I} \to \mathbb{R}$ provided that

$$\lim_{k\to\infty} \left[U\left(f, \mathbf{P}_k\right) - L\left(f, \mathbf{P}_k\right)\right] = 0.$$

Theorem (The Archimedes-Riemann Theorem): Let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function on the generalized rectangle $\mathbf{I}$. Then f is integrable on $\mathbf{I}$ if and only if there is an Archimedean sequence of partitions for $f : \mathbf{I} \to \mathbb{R}$. Moreover, for any such Archimedean sequence of partitions $\{\mathbf{P}_k\}$,

$$\lim_{k\to\infty} L\left(f, \mathbf{P}_k\right) = \int_{\mathbf{I}} f \quad \text{and} \quad \lim_{k\to\infty} U\left(f, \mathbf{P}_k\right) = \int_{\mathbf{I}} f. \tag{5.12}$$

The following theorem links this result with the univariate result in (2.128).

Theorem: Let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function on the generalized rectangle $\mathbf{I}$. Then the following two assertions are equivalent:

i. There is an Archimedean sequence of partitions for $f : \mathbf{I} \to \mathbb{R}$.

ii. For each $\epsilon > 0$ there is a partition $\mathbf{P}$ of $\mathbf{I}$ such that

$$U(f, \mathbf{P}) - L(f, \mathbf{P}) < \epsilon.$$

Proof: First we suppose that (i) holds. Let $\{\mathbf{P}_k\}$ be an Archimedean sequence of partitions for $f : \mathbf{I} \to \mathbb{R}$. To verify criterion (ii) we let ϵ be any positive number. By the definition of convergent sequence we can choose an index k such that $U\left(f, \mathbf{P}_k\right) - L\left(f, \mathbf{P}_k\right) < \epsilon$. Thus, setting $\mathbf{P} = \mathbf{P}_k$, we have $U(f, \mathbf{P}) - L(f, \mathbf{P}) < \epsilon$. Thus, criterion (ii) holds.

Now suppose that criterion (ii) holds. Let k be a natural number. Then, setting $\epsilon = 1/k$, according to (ii) there is a partition $\mathbf{P}$ such that $U(f, \mathbf{P}) - L(f, \mathbf{P}) < 1/k$. Choose such a partition and label it $\mathbf{P}_k$. This defines a sequence of partitions $\{\mathbf{P}_k\}$ of the generalized interval $\mathbf{I}$ that is Archimedean since

$$0 \leq \lim_{k\to\infty} \left[U\left(f, \mathbf{P}_k\right) - L\left(f, \mathbf{P}_k\right)\right] \leq \lim_{k\to\infty} 1/k = 0.$$

Here is the generalization of the *domain additivity*, or *additivity over partitions* from (2.138), which we state without proof: See, e.g., Fitzpatrick, p. 479.

Theorem (Additivity over Partitions): Let $f : \mathbf{I} \to \mathbb{R}$ be a bounded function on the generalized rectangle $\mathbf{I}$. Let $\mathbf{P}$ be a partition of $\mathbf{I}$. Then $f \in \mathcal{R}[\mathbf{I}]$ if and only if for each generalized rectangle $\mathbf{J}$ in $\mathbf{P}$, the restriction of f to $\mathbf{J}$, $f : \mathbf{J} \to \mathbb{R}$, is integrable: In this case,

$$\int_{\mathbf{I}} f = \sum_{\mathbf{J} \text{ in } \mathbf{P}} \int_{\mathbf{J}} f.$$

We also have the following results, which are clear analogs of their univariate counterparts.

Theorem (Monotonicity of the Integral): Suppose that the functions $f : \mathbf{I} \to \mathbb{R}$ and $g : \mathbf{I} \to \mathbb{R}$ are integrable, where $\mathbf{I}$ is a generalized rectangle in $\mathbb{R}^n$, and also suppose that $\forall \mathbf{x} \in \mathbf{I}$, $f(\mathbf{x}) \leq g(\mathbf{x})$. Then $\int_{\mathbf{I}} f \leq \int_{\mathbf{I}} g$.

Theorem (Linearity of the Integral): Suppose that the functions $f : \mathbf{I} \to \mathbb{R}$ and $g : \mathbf{I} \to \mathbb{R}$ are integrable, where $\mathbf{I}$ is a generalized rectangle in $\mathbb{R}^n$. Then for any two numbers α and β, the function $\alpha f + \beta g : \mathbf{I} \to \mathbb{R}$ also is integrable and

$$\int_{\mathbf{I}} [\alpha f + \beta g] = \alpha \int_{\mathbf{I}} f + \beta \int_{\mathbf{I}} g.$$

As hoped and expected, we have the generalization of (2.130), which we did prove.

Theorem: Let $f : \mathbf{I} \to \mathbb{R}$ be a continuous function on a generalized rectangle $\mathbf{I}$. Then f is integrable on $\mathbf{I}$. That is,

$$f \in \mathcal{C}^0 \implies f \in \mathcal{R}[\mathbf{I}]. \tag{5.13}$$

A proof in the general n case can be found in, e.g., Fitzpatrick, p. 484; and, somewhat more advanced, Terrell, A Passage to Modern Analysis, 2019, Prop. 12.5.2, p. 374.

For vector-valued functions $f : B \subset \mathbb{R}^n \to \mathbb{R}^m$, there is a natural extension of Riemann integration: It is just the elementwise integration of each component function. From Terrell, p. 364, we have:

Definition: Let $\mathbf{F} : B \to \mathbf{R}^m$ be a function bounded on the closed interval B in $\mathbf{R}^n$, and let us write $\mathbf{F} = (f_1, \ldots, f_m)$ where the f_j are the real valued component functions. We say that $\mathbf{F}$ is Riemann integrable on B if and only if each component function $f_j : B \to \mathbf{R}, 1 \leq j \leq m$, is Riemann integrable on B. Then the vector

$$\int_B \mathbf{F} = \left(\int_B f_1, \ldots, \int_B f_m \right)$$

is called the Riemann integral of $\mathbf{F}$ on B.

We close this subsection by mentioning a characterization of Riemann integrability that had to wait until the 20th century to be discovered (by Lebesgue). Some first-course analysis books cover the (long, without the use of measure theory) proof in the $n = 1$ case, such as the enjoyable presentation in Stoll (2021). We state that result in (2.131). The general n case is proved in Terrell, 2019, §12.5. The theorem refers to Lebesgue measure zero, the definition of which is not complicated and does not require a deep dive into measure theory. We discuss this in the subsequent subsection §5.2.

Theorem: Let B be a closed interval in $\mathbf{R}^n$. A bounded function $f : B \to \mathbf{R}$ is Riemann integrable on B if and only if the set of points where f is discontinuous has Lebesgue measure zero.

5.2 Bounded Sets, Jordan Measure, Volume Zero, and Lebesgue Measure Zero

A proper, rigorous development of the multivariate Riemann integral requires the concepts of Jordan measure (or Jordan content), volume, and volume zero. In this subsection, we only define these quantities and draw a contrast to the concept of Lebesgue measure zero.

5.2.1 Introduction and Useful Results

The Lebesgue integral has desirable features not possessed by the Riemann integral, and is the main reason that some analysis books cover the univariate Riemann integral, but then, for the multivariate case, skip the development of the Riemann integral, and go straight to the Lebesgue integral. The latter is an early 20th century development, and considered one of the most important advances in analysis. To understand it requires first learning what is called measure theory, this being a topic for a subsequent course. Below, we indicate two advantages of the notion of Lebesgue measure zero, as compared to that of Jordan volume zero, thus scratching the surface of why the Lebesgue integral is considered superior to the Riemann integral.

As a slight counterbalance, the formulation of the Lebesgue integral does not give rise to a method for numerically computing it, whereas the construction of the Riemann integral does. All algorithms for numeric integration (e.g., Simpson's rule, though there are far more sophisticated methods, and these are conveniently built in to numeric software, such as Matlab) are based on the Riemann formulation. The key result is that, if a function is Riemann integrable, then it is Lebesgue integrable, but not vice-versa. The Lebesgue integral formulation is of great use for theoretical reasons, but for computation, it is required that the function is also Riemann integrable.

We collect some useful definitions that we will require below.

Definition: An **open interval** in $\mathbb{R}^n$, $n \geq 2$, is a Cartesian product of n real intervals,

$$B = (a_1, b_1) \times \cdots \times (a_n, b_n)\,, \tag{5.14}$$

where $a_i < b_i$ for $1 \leq i \leq n$. A **closed interval** in $\mathbb{R}^n$, $n \geq 2$, has the form

$$B = [a_1, b_1] \times \cdots \times [a_n, b_n]\,, \tag{5.15}$$

where $a_i \leq b_i$ for $1 \leq i \leq n$. (Note that if $a_i < b_i$, $1 \leq i \leq n$, then the interior of a closed interval is the open interval having the same endpoints for each interval factor.) The volume of either of these types of intervals, described by the Cartesian product of real intervals, is defined to be $\nu(B) = \prod_{i=1}^{n} (b_i - a_i)$. In particular, note how the volumes of (5.14) and (5.15) are defined to be equal.

Definition: The volume of a union of finitely many intervals, any two of which intersect (if at all) only along boundary segments, is defined to be the sum (finite) of the volumes of the intervals.

This definition leads to (and is a special case of) result (5.20) below. Some authors refer to two intervals that intersect only along boundary segments as "nonoverlapping" (in the sense that, while they have points in common, in $\mathbb{R}^n$, this is a set of measure zero; recall (1.5) and also see below. Indeed, so does Terrell, on page 519, in his chapter on measure theory:

Two intervals in $\mathbb{R}^n$ (whether open, closed, or otherwise) are nonoverlapping if their interiors are disjoint, that is, they intersect only in some boundary points, if at all. Thus the intersection of the two intervals equals the intersection of their boundaries. Similarly, the intervals in an arbitrary collection of intervals are called nonoverlapping if any two of them are nonoverlapping.

The following presentation comes from Terrell, §12.2, §12.3, §12.4, and §12.5.

5.2.2 Bounded Sets, Jordan Measure, Volume Zero

The first task is to generalize the Riemann integral (5.11) to integration over other bounded sets. Let $S \subset \mathbb{R}^n$ be a bounded set, and $f : S \to \mathbb{R}$ a bounded function. We may extend f to all of $\mathbb{R}^n$ by defining

$$f_S(\mathbf{x}) = \begin{cases} f(\mathbf{x}), & \text{if } \mathbf{x} \in S, \\ 0, & \text{if } \mathbf{x} \notin S. \end{cases} \tag{5.16}$$

This is called the extension of f by zero. Let B be a closed interval in $\mathbb{R}^n$ that contains the bounded set S. We want to say that f is integrable on S if f_S is integrable on B, that is, if the integral $\int_B f_S$ exists. However, we have to show that the existence of the integral, and its value, is independent of the enclosing interval B.

Lemma: Let S be a bounded subset of $\mathbb{R}^n$ and $f : S \to \mathbb{R}$ a bounded function such that $\int_B f_S$ exists for some closed interval B containing S. Then

$$\int_B f_S = \int_{B'} f_S \tag{5.17}$$

for any other closed interval B' in $\mathbb{R}^n$ containing S.

We omit the proof, which can be found in Terrell, p. 366. The Lemma justifies the following definition.

Definition: If $S \subset \mathbb{R}^n$ is a bounded set and $f : S \to \mathbb{R}$ is a bounded function for which $\int_B f_S$ exists for some closed interval B containing S, then f is integrable on S and

$$\int_S f = \int_B f_S \tag{5.18}$$

is the integral of f on S. Thus the existence of $\int_S f$, and its value, are independent of the enclosing interval B.

Definition: If $A \subset \mathbb{R}^n$ is a bounded set, the **characteristic function** of A is the mapping $\chi_A : \mathbb{R}^n \to \mathbb{R}$ defined by

$$\chi_A(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x} \in A, \\ 0, & \text{if } \mathbf{x} \notin A. \end{cases} \tag{5.19}$$

Definition: The set A is **Jordan measurable**, or A **has volume**, if χ_A is integrable on A, that is, $\int_A \chi_A$ exists.

Definition: The **volume** of A, denoted $\nu(A)$, is defined by

$$\nu(A) = \int_A \chi_A.$$

Definition: For open interval S and its closure $\bar{S}$,

$$S = (a_1, b_1) \times \cdots \times (a_n, b_n)\,, \qquad \bar{S} = [a_1, b_1] \times \cdots \times [a_n, b_n]\,,$$

their volumes equal $\prod_{i=1}^{n} (b_i - a_i)$ by axiom. We define, for consistency,

$$\int_S \chi_S = \int_{\bar{S}} \chi_{\bar{S}} = \prod_{i=1}^{n} (b_i - a_i)\,.$$

As important special cases: For a subset A of the two-dimensional plane, the volume is the area of the region, and this area is numerically equal to the (three-dimensional) volume of the solid lying between the graph of χ_A and the region A in the plane. For an interval $A = [a, b]$ of real numbers, the volume is the length of the interval, and this length is numerically the same as the area of the region between the graph of $\chi_{[a,b]}$ and the interval $A = [a, b]$ on the real line.

Definition: The volume of A, when it exists, is also called the **Jordan measure**, or **Jordan content**, of A.

Definition: A set A with volume such that $\nu(A) = 0$ is said to have **volume zero**. The concept of volume zero is also called **Jordan measure zero** or **Jordan content zero**.

The following results on Jordan measure (from Terrell, p. 379) parallel fundamental results for Lebesgue measure (which we do not state here):

Theorem: Let S_1 and S_2 be subsets of $\mathbb{R}^n$ that have volume. Then:

1. $S_1 \cup S_2$ and $S_1 \cap S_2$ have volume, and

$$\nu\,(S_1 \cup S_2) = \nu\,(S_1) + \nu\,(S_2) - \nu\,(S_1 \cap S_2)\,.$$

2. If Int $S_1 \cap$ Int S_2 is the empty set, then

$$\nu\,(S_1 \cup S_2) = \nu\,(S_1) + \nu\,(S_2)\,. \tag{5.20}$$

3. If $S_1 \subseteq S_2$, then $S_2 - S_1 = S_2 \cap S_1^c$ has volume and

$$\nu\,(S_2 \cap S_1^c) = \nu\,(S_2) - \nu\,(S_1)\,.$$

4. If $S_1 \subseteq S_2$, then

$$\nu\,(S_1) \leq \nu\,(S_2)\,.$$

Proposition: From the definition of integrability of χ_A, a set A has volume zero if and only if, for every $\epsilon > 0$, there is a finite collection of closed intervals $C_1, \ldots, C_N$ such that

$$A \subseteq \bigcup_{i=1}^{N} C_i \quad \text{and} \quad \sum_{i=1}^{N} \nu\,(C_i) < \epsilon. \tag{5.21}$$

Author Terrell assigns this as an exercise (12.3.2), with the hint (and note he refers to boxes in $\mathbb{R}^n$ as intervals): For each implication, think of the intervals C_i as intervals of a partition P, involved in defining an upper sum $U\,(\chi_A, P)$ for that partition.

($\Rightarrow$) We have $\nu(A) = \int_A \chi_A = 0$, i.e., by the definition of volume, χ_A is integrable. From (5.11), this means that χ_A is bounded, which is trivially true; but also that A is either a generalized rectangle, which from (5.1) and (5.2) means it is bounded; or, from definition (5.16), a more general bounded set. From (5.18), we can take A to be its closure, $\bar{A}$, which is necessarily closed and bounded. Note it is the smallest closed set containing A. Being bounded, there exists a generalized rectangle $\mathbf{I}$ that covers A, and from (5.18), $\int_A \chi_A = \int_{\mathbf{I}} \chi'_A$, where χ'_A is the extension of function χ_A defined in (5.16).

Let $\mathbf{P}_1 = \mathbf{I}$; and let $\mathbf{P}_k$ be a sequence of partitions of A with $\mathbf{P}_k \subset \mathbf{P}_{k+1}$, i.e., the latter is a refinement of the former. From (5.12), $\lim_{k\to\infty} U(\chi_A, \mathbf{P}_k) = \int_A \chi_A = 0$, and we know the rhs integral exists by assumption, while for any $k \in \mathbb{N}$, the lhs exists. From the definition of limit, for each $j \in \mathbb{N}$, $\exists K_j \in \mathbb{N}$ such that, for $k \geq K_j$, $U(\chi_A, \mathbf{P}_k) < 1/j$. For $j > 1/\epsilon$, the finite partition $\mathbf{P}_m$, $m = K_j$, satisfies $U(\chi_A, \mathbf{P}_m) < \epsilon$. The value N in (5.21) is the number of generalized rectangles in $\mathbf{P}_m$.

($\Leftarrow$) We are given that, $\forall \epsilon > 0$, $\exists \{C_i\}_{i=1}^N$, C_i closed, such that (5.21) holds. We require use of (5.18) to handle $A \subseteq \cup_{i=1}^N C_i$; and, if we assume the C_i are nonoverlapping, (5.20) to justify the equality $\nu\left(\cup_{i=1}^N C_i\right) = \sum_{i=1}^N \nu(C_i)$, so $\nu\left(\cup_{i=1}^N C_i\right) < \epsilon$. We can take the $\{C_i\}$ to be nonoverlapping, and let $\mathbf{P}_k = \{C_i\}$. Then, by the definition of χ_A in (5.19) (namely that it is constant on A with value 1, and zero otherwise), we have $0 \leq L(\chi_A, \mathbf{P}_k) \leq U(\chi_A, \mathbf{P}_k) < \epsilon$. Thus, from (5.12) and the Squeeze Theorem (2.3), $\int_A \chi_A$ exists and equals zero, i.e., A has volume zero.

Having now secured the solutions manual to Terrell's book, we can compare my above attempt to his proof.

Proof (Terrell): Suppose A has volume zero, that is, $\nu(A) = \int_A \chi_A = 0$, and B is a closed interval containing A. Then for every $\epsilon > 0$, there is a partition P of B such that $U(\chi_A, P) < \epsilon$. Let the listing $S_1, \ldots, S_N$ be an enumeration of the intervals of the partition P. Then $A \subseteq \bigcup_{i=1}^N S_i$ and $\sum_{i=1}^N \nu(S_i) = U(\chi_A, P) < \epsilon$.

Suppose that for every $\epsilon > 0$ there is a finite collection of closed intervals $S_1, \ldots, S_N$ such that $A \subseteq \bigcup_{i=1}^N S_i$ and $\sum_{i=1}^N \nu(S_i) < \epsilon$. Let B be any closed interval that contains the union of the S_i. The intervals S_i may intersect or not, but in any case, there is a partition P of B that includes all the lattice points defined by the endpoints of the $S_i, 1 \leq i \leq N$. Then we have $U(\chi_A, P) \leq \sum_{i=1}^N \nu(S_i) < \epsilon$. Since $L(\chi_A, P) = 0$, this shows, by the Riemann criterion, that χ_A is integrable and $\nu(A) = \int_A \chi_A = 0$.

One of the weaknesses of the volume concept is that it does not apply to unbounded sets. Another weakness is that a countable union of sets having volume is not necessarily a set having volume, even in some cases where we think it probably should be. As an example, on the real line, the open set $\bigcup_{k=1}^\infty \left(k - 1/2^k, k + 1/2^k\right)$ has what we call finite total length, given by

$$\sum_{k=1}^\infty \frac{2}{2^k} = \sum_{k=1}^\infty \frac{1}{2^{k-1}} = 2,$$

but it does not have volume since it is an unbounded set.

Now consider the next example: Any single point, that is, a singleton set $\{\mathbf{x}\}$, has volume zero, since it can be covered by a single closed interval of arbitrarily small volume. On the real line, consider the rational numbers in $[0, 1]$, that is, $S = \mathbf{Q} \cap [0, 1]$. Then S is bounded, and it is the union of countably many (singleton) sets of volume zero, but S does not have

volume, much less volume zero, since χ_S is not integrable. There are similar examples in the plane and in higher dimensions. For example, the rational points (the points with rational coordinates) in the unit square in the plane, $\mathbf{Q} \times \mathbf{Q} \cap [0,1] \times [0,1]$, is a countable union of sets, each with volume zero, but it does not have volume, since its characteristic function is not integrable.

We have seen that there are open sets that do not have volume. Since open sets play a fundamental role in analysis, this must be seen as a weakness in the theory of Jordan measure we are discussing, and the weakness is tied to the Riemann integral concept through the above definition of volume, which relies on the existence of the Riemann integral. The central issue that prevents some bounded sets S from having volume is that the boundary ∂S may be too complicated to allow integrability of the characteristic function χ_S. This issue about the boundary ∂S is discussed subsequently.

5.2.3 Lebesgue Measure Zero

We defined Lebesgue measure zero in the univariate case in (1.5). Its extension to $\mathbb{R}^n$ is very natural.

Definition: Let $S \subset \mathbb{R}^n$, bounded or unbounded. We say that S has n-dimensional Lebesgue measure zero (or simply measure zero) if, for every $\epsilon > 0$, there is a sequence of open intervals, J_i, in $\mathbb{R}^n$ such that $S \subseteq \bigcup_i J_i$ and

$$\sum_i \nu(J_i) < \epsilon.$$

The concepts of measure zero and volume zero depend on the dimension, and one can write $m_n(S) = 0$ and $\nu_n(S) = 0$ to indicate n-dimensional Lebesgue measure zero and n-dimensional volume zero, respectively, if needed.

Example 5.1 *Let S be the set of rational numbers in the unit interval, $S = \mathbf{Q} \cap [0,1]$. Then S has Lebesgue measure zero. We enumerate these rationals by the listing $\{r_1, r_2, r_3, \ldots\}$, and then cover the numbers individually by open intervals whose lengths sum to less than a given $\epsilon > 0$. For example, cover r_1 by an open interval of length $\epsilon/2$, r_2 by an open interval of length $\epsilon/2^2$; in general, cover r_k by an open interval of length $\epsilon/2^k$. Then the countable collection of these open intervals covers S and has total length less than $\sum_{k=1}^{\infty} \epsilon/2^k = \epsilon$. Therefore S has Lebesgue measure zero.* ■

Example 5.2 *The set $S = \{(x,0) : 0 \le x \le 1\}$ has 2-dimensional Lebesgue measure zero. To verify this, observe that S can be covered by the single closed interval $[0,1] \times [0,\delta]$ for any $\delta > 0$. Since this interval has volume δ, we conclude that S has volume zero, and hence S has measure zero. Alternatively, given $0 < \epsilon < 1$, S can be covered by a single open interval, for example,*

$$R = \left\{(x,y) : -\frac{\epsilon}{4} < x < 1 + \frac{\epsilon}{4}, -\frac{\epsilon}{4} < y < \frac{\epsilon}{4}\right\}$$

which has volume $\nu(R) = (1 + \epsilon/2)(\epsilon/2) < \epsilon$. Therefore S has measure zero. ■

Theorem: If S has n-dimensional volume zero, then it has n-dimensional Lebesgue measure zero.

Proof: If S has volume zero, then, from (5.21), for any $\epsilon > 0$, S can be covered by a finite collection of closed intervals $I_i, 1 \leq i \leq N$, such that $\sum_{i=1}^{N} \nu(I_i) < \epsilon/2$. For each i, we can cover I_i with an open interval J_i of volume $\nu(I_i) + \epsilon/2^{i+1}$, and $\sum_{i=1}^{N} \nu(J_i) = \sum_{i=1}^{N} \nu(I_i) + \sum_{i=1}^{N} \epsilon/2^{i+1} < \epsilon/2 + \epsilon/2 = \epsilon$. Since countable means finite or countably infinite, S has Lebesgue measure zero.

On the other hand, there are sets having Lebesgue measure zero that do not have volume, as we see in the next example.

Example 5.3 *Let S be the set of points in the unit square $[0,1] \times [0,1]$ having rational coordinates, that is,*

$$S = \mathbf{Q} \times \mathbf{Q} \cap [0,1] \times [0,1].$$

Then S has Lebesgue measure zero, since S is countable. However, S does not have volume, because the characteristic function of S is not integrable. ■

We have seen that volume zero implies Lebesgue measure zero; however, the converse does not generally hold. An exception is described in the next proposition.

Proposition: A compact set in $\mathbb{R}^n$ that has Lebesgue measure zero also has volume zero.

Proof: Suppose $A \subset \mathbb{R}^n$ is compact (that is, closed and bounded) and has Lebesgue measure zero. Let $\epsilon > 0$. Since A has Lebesgue measure zero, there is a sequence of open intervals, J_i, in $\mathbb{R}^n$ such that $A \subseteq \bigcup_i J_i$ and $\sum_i \nu(J_i) < \epsilon$. Since A is compact, there is a finite subcover $\{J_{i_1}, J_{i_2}, \ldots, J_{i_M}\}$ of A. By taking the closure of each of these M open intervals, we have the collection $\left\{\bar{J}_{i_1}, \bar{J}_{i_2}, \ldots, \bar{J}_{i_M}\right\}$ of closed intervals, which covers A, and

$$\sum_{j=1}^{M} \nu\left(\bar{J}_{i_j}\right) \leq \sum_{i} \nu(J_i) < \epsilon.$$

This argument holds for every $\epsilon > 0$, and therefore, again from (5.21), A has volume zero.

Observe that, if $J_1 \times \cdots \times J_n$ is an interval in $\mathbb{R}^n$, then its boundary is given by

$$\bigcup_{k=1}^{n} J_1 \times \cdots \times J_{k-1} \times (\partial J_k) \times J_{k+1} \times \cdots \times J_n.$$

It is not difficult to see that the boundary of an interval in $\mathbb{R}^n$ has volume zero, and thus the boundary has n-dimensional Lebesgue measure zero. On the other hand, it is not difficult to directly see that the boundary of an interval in $\mathbb{R}^n$ has Lebesgue measure zero, and that the boundary is compact (since it is closed and bounded); hence, the boundary of an interval has volume zero by the proposition.

It follows directly from the definition that every subset of a set of measure zero has measure zero. In particular, since the empty set is a subset of every set, it is covered by a single interval of arbitrarily small volume, and hence has Lebesgue measure zero. By a similar argument, any singleton set $\{\mathbf{x}\}$ has measure zero, and thus every finite set in $\mathbb{R}^n$ has Lebesgue measure zero.

Example 5.4 *Let us show that the graph of a continuous function $f : [a,b] \to \mathbf{R}$ has 2-dimensional Lebesgue measure zero. It suffices to show that the graph has 2-dimensional volume zero. Let $G = \{(x, f(x)) : x \in [a,b]\}$ be the graph.*

From (2.27) or (2.31), f is uniformly continuous on $[a,b]$. This means that, for every $\epsilon > 0$, there is a $\delta > 0$ such that $|x_1 - x_2| < \delta$ implies $|f(x_1) - f(x_2)| < \epsilon/(b-a)$. Thus, for every $\epsilon > 0$, we can find a finite cover of G by closed rectangles having height $\epsilon/(b-a)$ and nonoverlapping interiors. Thus, G is covered by finitely many closed intervals in $\mathbf{R}^2$ whose total volume is less than or equal to ϵ, so $\nu(G) = 0$. ■

Remark: Despite the result of this last example, a continuous image of a set with n-dimensional volume zero need not have n-dimensional volume zero. This fact is demonstrated by the existence of so-called space-filling curves.

Example 5.5 *It is not true that the boundary of every set $E \subseteq \mathbb{R}$ has measure zero. For example, the set of rationals $\mathbb{Q}$ has measure zero, but its boundary is $\partial\mathbb{Q} = \mathbb{R}$, whose measure is infinite.* ■

An advantage of the concept of measure zero over that of volume zero is that the union of a countable infinity of sets, each having measure zero, is also a set of measure zero:

Theorem: A countably infinite union of sets of n-dimensional Lebesgue measure zero is a set of n-dimensional Lebesgue measure zero.

The proof of this result invokes a result involving a monotone increasing sequence and a subsequence thereof, namely:

> If (b_k) is an increasing sequence and if some subsequence (b_{n_k}) of (b_k) converges and $\lim_{k\to\infty} b_{n_k} = b$, then (b_k) itself converges to the same limit, $\lim_{k\to\infty} b_k = b$. See, e.g., Terrell, p. 42 for proof.

Proof: Let $\{E_k\}$ be a countable collection of subsets $E_k \subset \mathbb{R}^n$, each having Lebesgue measure zero. Given $\epsilon > 0$, there exists a doubly indexed collection of open intervals $\left\{J_j^k\right\}$ in $\mathbb{R}^n$ such that, for each k,

$$\bigcup_j J_j^k \supset E_k \quad \text{and} \quad \sum_j \nu\left(J_j^k\right) < \frac{\epsilon}{2^{k+1}}.$$

Thus, $\bigcup_{k,j} J_j^k$ covers $\bigcup_k E_k$. The problem now is to arrange this doubly indexed collection of intervals into a sequence and then sum the volumes according to that definite sequence.

We arrange the volumes $\nu\left(J_j^k\right)$ of these intervals in a matrix with $\nu\left(J_1^1\right)$ as the upper left entry and, using k as the row index, j as the column index, we can list the volumes, by tracing the diagonals of slope one in our matrix, starting from the upper left entry, in the order

$$\nu\left(J_1^1\right), \quad \nu\left(J_1^2\right), \nu\left(J_2^1\right), \quad \nu\left(J_1^3\right), \nu\left(J_2^2\right), \nu\left(J_3^1\right), \quad \nu\left(J_1^4\right), \nu\left(J_2^3\right), \nu\left(J_3^2\right), \nu\left(J_4^1\right), \tag{5.22}$$

and so on. Let σ_n denote the nth partial sum based on this ordering of the volumes; thus, $\sigma_1 = \nu\left(J_1^1\right)$, $\sigma_2 = \nu\left(J_1^1\right) + \nu\left(J_1^2\right)$, $\sigma_3 = \nu\left(J_1^1\right) + \nu\left(J_1^2\right) + \nu\left(J_2^1\right)$, etc.. With the diagonals of slope one in our matrix in view, we can form the triangular partial sums

$$s_n = \sum_{k+j\le n} \nu\left(J_j^k\right)$$

and notice that we have

$$s_1 = \sigma_1, \quad s_2 = \sigma_3, \quad s_3 = \sigma_6, \quad \ldots, \quad s_n = \sigma_{(n(n+1))/2}, \quad \ldots.$$

Thus (σ_n) is an increasing sequence with subsequence (s_n). By (5.22), we may write

$$s_n = \sigma_{(n(n+1))/2} = \sum_{k=1}^{n} \left(\nu\left(J_1^k\right) + \nu\left(J_2^{k-1}\right) + \cdots + \nu\left(J_k^1\right) \right).$$

Since $\nu\left(J_j^k\right) \geq 0$ for all $k, j \in \mathbf{N}$, we have

$$\begin{aligned} s_n = \sigma_{(n(n+1))/2} &= \sum_{k=1}^{n} \left(\nu\left(J_1^k\right) + \nu\left(J_2^{k-1}\right) + \cdots + \nu\left(J_k^1\right) \right) \leq \sum_{k=1}^{n} \left(\sum_{j=1}^{n} \nu\left(J_j^k\right) \right) \\ &\leq \sum_{k=1}^{n} \left(\sum_{j=1}^{\infty} \nu\left(J_j^k\right) \right) < \sum_{k=1}^{n} \frac{\epsilon}{2^{k+1}} < \frac{\epsilon}{2} < \epsilon. \end{aligned}$$

Therefore the sequence $(s_n) = \left(\sigma_{(n(n+1))/2}\right)$ is increasing and bounded above by ϵ, and it converges to a limit $s < \epsilon$. Since it is a subsequence of the increasing sequence (σ_n), we conclude from the above theorem that (σ_n) itself converges to the same limit, hence $\lim_{n\to\infty} \sigma_n = s < \epsilon$. We conclude that the sum of the volumes of the intervals J_j^k, listed in (5.22), is less than ϵ. Since ϵ is arbitrary, this shows that $\bigcup_k E_k$ has Lebesgue measure zero.

Example 5.6 *The real numbers of the form $a + b\sqrt{2}, a, b \in \mathbb{Q}$, are all irrational. The set of such numbers is a countable union of countable sets, and hence has measure zero.* ■

We noted that the real interval [0,1], considered as a subset of the plane, has 2-dimensional Lebesgue measure zero. Let us show that the entire real line, considered as a subset of the plane, has measure zero.

Example 5.7 *The entire real line, considered as a subset of the plane, has measure zero. The proof depends on showing that increasingly larger chunks of the embedded line can be covered by smaller and smaller 2-dimensional interval volumes. Given $0 < \epsilon < 1$, we must find open intervals J_j in $\mathbf{R}^2$ such that $\sum_j \nu(J_j) < \epsilon$. For example, we may choose*

$$J_j = \left(-\frac{j}{2}, \frac{j}{2} \right) \times \left(-\frac{\epsilon}{2j\,(2^j)}, \frac{\epsilon}{2j\,(2^j)} \right).$$

The 2-dimensional volume of J_j is $\nu(J_j) = \epsilon/2^j$, and thus $\sum_j \nu(J_j) < \epsilon$. Moreover, the entire real line, considered as a subset of the plane (i.e., the set $\{(x,0) : x \in \mathbf{R}\}$), is contained in the union of the J_j. Thus, the embedded real line has measure zero in $\mathbb{R}^2$. ■

We state some further fundamental results, without proof, from Terrell, §12.5. The first generalizes the $n = 1$ case stated in (2.131) (where other references for the proof can be found):

Proposition: Let B be a closed interval in $\mathbb{R}^n$. If $f : B \to \mathbb{R}$ is a bounded function and the set D of discontinuities of f has Lebesgue measure zero, then $f \in \mathcal{R}[B]$.

Proposition: Let B be a closed interval in $\mathbb{R}^n$. If $f : B \to \mathbb{R}$ is Riemann integrable on B, then the set D of discontinuities of f has Lebesgue measure zero.

These two previous propositions imply:

Theorem: Let B be a closed interval in $\mathbb{R}^n$, and $f : B \to \mathbb{R}$ bounded. $f \in \mathcal{R}[B]$ if and only if the set of points where f is discontinuous has Lebesgue measure zero.

Corollary: Let S be a bounded set in $\mathbb{R}^n$ and B any closed interval containing S. A bounded function $f : S \to \mathbb{R}$ is integrable on S if and only if the set of discontinuities of f_S, the extension of f by zero to B, has Lebesgue measure zero;

Corollary: A bounded set S in $\mathbb{R}^n$ has volume if and only if ∂S has Lebesgue measure zero.

Definition: If a property or statement involving points of B holds for all points except on a subset of B of Lebesgue measure zero, we say that the property holds almost everywhere (a.e.) in B.

Corollary: Let S be a bounded set that has volume. A bounded function $f : S \to \mathbb{R}$ is integrable on S if and only if f is continuous a.e. in the interior of S.

Example 5.8 *(Heil, Measure Theory for Scientists and Engineers, p. 97)*
Define $f : [0, \infty) \to [0, \infty]$ *by*

$$f(t) = \begin{cases} 1/t, & \text{if } t > 0, \\ \infty, & \text{if } t = 0. \end{cases}$$

This function takes finite values at all but a single point. Hence the set

$$Z = \{f = \pm\infty\} = \{t \in [0, \infty) : f(t) = \pm\infty\}$$

where f *is not finite has measure zero, so we say that*

$$f(t) \text{ is finite for almost every } t \in [0, \infty),$$

or simply that f *is finite a.e..*

Every bounded function is certainly finite a.e., but the function f *is an example of function that is finite almost everywhere but not bounded.*

If $f : E \to [-\infty, \infty]$ *is a generic extended real-valued function, then* f *is bounded implies* f *is finite a.e.; but* f *is finite a.e. does not imply* f *is bounded.* ■

5.3 Exchange of Derivative and Integral

Throughout this subsection, let $a_1, a_2, b_1, b_2 \in \mathbb{R}$ with $a_1 < b_1$, $a_2 < b_2$, and let $D := [a_1, b_1] \times [a_2, b_2]$ be a closed rectangle in $\mathbb{R}^2$. To get to the result we want, we first prove some basic results. We will also need these in the next subsection on Fubini's theorem.

Theorem: If $f : D \to \mathbb{R}$ is continuous, then so are

$$\phi_1(x) := \int_{a_1}^{b_1} f(t,x)\,dt \quad \text{and} \quad \phi_2(t) := \int_{a_2}^{b_2} f(t,x)\,dx. \tag{5.23}$$

We prove a more general statement just below. For now, recall the FTC (ii, a) in (2.145):

For $f \in \mathcal{R}[a,b]$, $F(x) = \int_a^x f$, and $x \in I = [a,b]$, we have $F \in \mathcal{C}^0[a,b]$.

Recall from (2.130) that continuity implies integrability, giving an important special case of FTC (ii, a). Observe how result (5.23) generalizes this.

The second statement in (5.23) is a special case of the following result, also required below.

Theorem: If $f : D \to \mathbb{R}$ is continuous,

$$\forall (t,x) \in D, \quad \psi(t,x) := \int_{a_2}^{x} f(t,u)\,du \quad \textbf{is continuous}. \tag{5.24}$$

Proof: Recalling (4.9), we need to show that, $\forall (t,x) \in D$ and any given $\epsilon > 0$,

$$\exists \delta \text{ such that } (t_0,x_0) \in \mathcal{B}_\delta((t,x)) \cap D \implies |\psi(t,x) - \psi(t_0,x_0)| < \epsilon. \tag{5.25}$$

Using domain additivity (2.138), and defining A and B as the indicated two integrals,

$$\begin{aligned} \psi(t,x) - \psi(t_0,x_0) &= \int_{a_2}^{x} f(t,u)\,du - \int_{a_2}^{x_0} f(t_0,u)\,du \\ &= \int_{a_2}^{x_0} [f(t,u) - f(t_0,u)]\,du + \int_{x_0}^{x} f(t,u)\,du \\ &=: A + B. \end{aligned} \tag{5.26}$$

As D is closed and bounded and f is continuous on D, from (2.28), f is bounded by some number, say K; and, from (2.27), is uniformly continuous on D.

Let $\epsilon > 0$. To bound the (absolute value of the) second integral in (5.26), choose x_0 such that $|x - x_0| < \epsilon/K$.

For the first integral, as f is uniformly continuous, there exists δ_1 such that, whenever $|t - t_0| < \delta_1$, $|f(t,u) - f(t_0,u)| < \epsilon$.

Let $\delta = \min(\epsilon/K, \delta_1)$. Then for $|x - x_0| < \delta$ and $|t - t_0| < \delta$, (5.26) is such that, from the triangle inequality,

$$\psi(t,x) - \psi(t_0,x_0) \le |\psi(t,x) - \psi(t_0,x_0)| \le |A| + |B| < \epsilon\,|x_0 - a_2| + \epsilon.$$

Observe that this is equivalent to (5.25), thus proving ψ is continuous.

Our goal is to know the conditions under which differentiation and integration can be exchanged. Let $f : D \to \mathbb{R}$ and $D_2 f$ be continuous on D, where, from (4.20),

$$D_2 f(t,x) = \lim_{h \to 0} \frac{f(t, x+h) - f(t,x)}{h}.$$

Theorem: Function $g(x) := \int_{a_1}^{b_1} f(t,x)\, dt$ is differentiable, and

$$g'(x) = \int_{a_1}^{b_1} D_2 f(t,x)\, dt. \tag{5.27}$$

Proof: As $D_2 f$ is continuous on D, (2.130) implies that $\int_{a_1}^{b_1} D_2 f(t,x)\, dt$ exists, so if (5.27) is true, then g is differentiable. To show (5.27), as in Lang (1997, p. 276), write

$$\frac{g(x+h) - g(x)}{h} - \int_{a_1}^{b_1} D_2 f(t,x)\, dt = \int_{a_1}^{b_1} \left[\frac{f(t,x+h) - f(t,x)}{h} - D_2 f(t,x) \right] dt.$$

By the MVT (2.63), for each t there exists a number $c_{t,h}$ between x and $x+h$ such that

$$\frac{f(t,x+h) - f(t,x)}{h} = D_2 f(t, c_{t,h}).$$

From (2.31), $D_2 f$ is uniformly continuous on the closed, bounded interval D, so

$$\left| \frac{f(t,x+h) - f(t,x)}{h} - D_2 f(t,x) \right| = |D_2 f(t,c_{t,h}) - D_2 f(t,x)| < \frac{\epsilon}{b_1 - a_1},$$

where $\epsilon > 0$, whenever h is sufficiently small.

NOTE: A sufficient condition for result (5.27) to be true when $b_1 = \infty$ is that f and $D_2 f$ are absolutely convergent. That is, for $D := [a_1, \infty) \times [a_2, b_2]$, $a_2 < b_2$, (5.27) holds if there are nonnegative functions $\phi(t)$ and $\psi(t)$ such that $|f(t,x)| \le \phi(t)$ and $|D_2 f(t,x)| \le \psi(t)$ for all $t, x \in D$, and $\int_{a_1}^{\infty} \phi$ and $\int_{a_1}^{\infty} \psi$ converge. See, e.g., Lang (1997, p. 337) for proof.

Example 5.9 *To calculate the derivative at zero of the function* $f(t) = \int_{-1}^{1} \sin(ts) e^{s+t} ds$, *differentiate under the integral sign, giving*

$$f'(t) = \int_{-1}^{1} \left(s \cos(ts) e^{s+t} + \sin(ts) e^{s+t} \right) ds,$$

so that

$$f'(0) = \int_{-1}^{1} \left(s \cos(0s) e^{s+0} + \sin(0s) e^{s+0} \right) ds = \int_{-1}^{1} (s e^s + 0)\, ds = (s-1) e^s \Big|_{-1}^{1} = 2e^{-1}.$$

This method is quite straightforward (at least for $t = 0$*) and obviates the need for a direct calculation of the complicated integral expression of* f. ■

5.4 Fubini's Theorem

We concentrate on stating and proving the two dimensional case, from which the full generalization becomes plausible.

Theorem: Let $a_1, a_2, b_1, b_2 \in \mathbb{R}$ with $a_1 < b_1$, $a_2 < b_2$, and let $D := [a_1, b_1] \times [a_2, b_2]$. Let $f : D \to \mathbb{R}$ be a continuous function. From (5.13), f is Riemann integrable on the set D, and we can use (a special case of the more general, given below) *Fubini's theorem*, due to Guido Fubini (1879-1943), to calculate its integral:

$$\int_D f(\mathbf{x})\, d\mathbf{x} = \int_{a_2}^{b_2} \left[\int_{a_1}^{b_1} f(x_1, x_2) dx_1 \right] dx_2 = \int_{a_1}^{b_1} \left[\int_{a_2}^{b_2} f(x_1, x_2) dx_2 \right] dx_1. \tag{5.28}$$

Observe that (5.28) is a set of nested *univariate* Riemann integrals. This can be extended in an obvious way to the n-dimensional case with $\mathbf{x} = (x_1, \ldots, x_n)$. Fubini's theorem holds whenever f is Riemann integrable; in particular, when $f : A \subset \mathbb{R}^n \to \mathbb{R}$ is continuous and A is closed and bounded.

Proof: As in Lang (1997, p. 277), we wish to show that

$$\int_{a_2}^{b_2} \left[\int_{a_1}^{b_1} f(t, x)\, dt \right] dx = \int_{a_1}^{b_1} \left[\int_{a_2}^{b_2} f(t, x)\, dx \right] dt.$$

Let $\psi(t, x) = \int_{a_2}^{x} f(t, u)\, du$, so that $D_2\psi(t, x) = f(t, x)$ from the FTC (2.146), and ψ is continuous from (5.24). We can now apply (5.27) to ψ and $D_2\psi = f$. Let $g(x) = \int_{a_1}^{b_1} \psi(t, x)\, dt$. Then

$$g'(x) = \int_{a_1}^{b_1} D_2\psi(t, x)\, dt = \int_{a_1}^{b_1} f(t, x)\, dt,$$

and, from the FTC (2.143),

$$g(b_2) - g(a_2) = \int_{a_2}^{b_2} g'(x)\, dx = \int_{a_2}^{b_2} \left[\int_{a_1}^{b_1} f(t, x)\, dt \right] dx.$$

On the other hand, from FTC (i) (2.143), $\int_{a_2}^{b_2} f(t, u)\, du = \psi(t, b_2) - \psi(t, a_2)$, so that

$$g(b_2) - g(a_2) = \int_{a_1}^{b_1} \psi(t, b_2)\, dt - \int_{a_1}^{b_1} \psi(t, a_2)\, dt = \int_{a_1}^{b_1} \left[\int_{a_2}^{b_2} f(t, u)\, du \right] dt,$$

and the theorem is proved.

The above proof, from Lang's book, is the most elegant, easy, and short I have seen. The reader can compare the other approach, as taken in Fitzpatrick, §19.1, and Terrell, §12.7. All three books provide further extensions and examples, and are worth reading. We look at two important extensions below.

Example 5.10 *(Shimamoto, Example 5.3) Consider the iterated integral* $\int_0^2 \left(\int_{x^2}^4 x^3 e^{y^3} dy \right) dx$. *We wish to sketch the domain of integration D in the xy-plane; and then evaluate the integral.*

The domain of integration can be reconstructed from the endpoints of the integrals. The outermost integral says that x goes from $x = 0$ to $x = 2$. Geometrically, we are integrating

areas of cross-sections that are perpendicular to the x-axis. Then the inner integral says that, for each x, y goes from $y = x^2$ to $y = 4$. The left panel of Figure 36 exhibits the relevant input. Hence D is described by the conditions:

$$D = \left\{(x, y) \in \mathbb{R}^2 : 0 \le x \le 2, x^2 \le y \le 4\right\}.$$

This is the region in the first quadrant bounded by the parabola $y = x^2$, the line $y = 4$, and the y-axis.

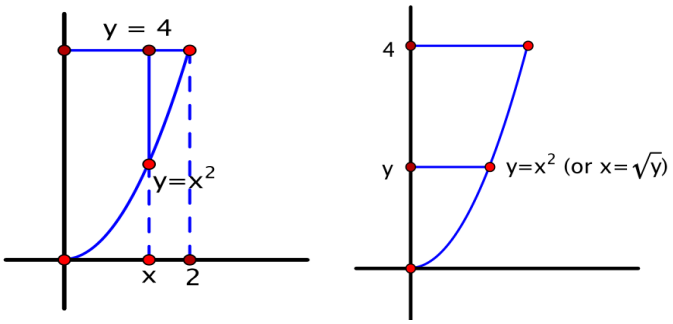


Figure 36: From Shimamoto, Multivariate Calculus, page 126.

To evaluate the integral as presented, we would antidifferentiate first with respect to y, treating x as constant:

$$\int_0^2 \left(\int_{x^2}^4 x^3 e^{y^3} dy\right) dx = \int_0^2 x^3 \left(\int_{x^2}^4 e^{y^3} dy\right) dx.$$

The innermost antiderivative looks hard. So, having nothing better to do, we try switching the order of antidifferentiation. Using cross-sections perpendicular to the y-axis, we see from the description of D that y goes from $y = 0$ to $y = 4$, and, for each y, x goes from $x = 0$ to $x = \sqrt{y}$. The thinking behind the switched order is illustrated in the right panel of Figure 36. Therefore,

$$\begin{aligned}
\int_0^2 \left(\int_{x^2}^4 x^3 e^{y^3} dy\right) dx &= \int_0^4 \left(\int_0^{\sqrt{y}} x^3 e^{y^3} dx\right) dy = \int_0^4 \left(\frac{1}{4} x^4 e^{y^3}\Big|_{x=0}^{x=\sqrt{y}}\right) dy \\
&= \int_0^4 \left(\frac{1}{4} y^2 e^{y^3} - 0\right) dy \quad \left(\text{let } u = y^3, du = 3y^2 dy\right) \\
&= \frac{1}{4} \cdot \frac{1}{3} e^{y^3}\Big|_0^4 = \frac{1}{12}\left(e^{64} - 1\right). \qquad \blacksquare
\end{aligned}$$

Example 5.11 *The above proof of Fubini's theorem used the interchange of derivative and integral result (5.27). It is instructive to go the other way, proving (5.27) with the use of Fubini's theorem. With $D = [a_1, b_1] \times [a_2, b_2]$ and $f : D \to \mathbb{R}$ and $D_2 f$ continuous functions, we wish to show that, for $t \in (a_2, b_2)$,*

$$\frac{d}{dt} \int_{a_1}^{b_1} f(x_1, t)\, dx_1 = \int_{a_1}^{b_1} D_2 f(x_1, t)\, dx_1. \tag{5.29}$$

Define $h : [a_2, b_2] \to \mathbb{R}$ as $h(x_2) := \int_{a_1}^{b_1} D_2 f(x_1, x_2)\, dx_1$, so that $h(t)$ is the rhs of (5.29), and, from FTC (ii, b) (2.146),

$$\frac{d}{dt}\int_{a_2}^{t} h(x_2)\,dx_2 = h(t). \tag{5.30}$$

As $D_2 f$ is continuous, it follows from (5.23) that h is continuous on $[a_2, b_2]$. Choosing an arbitrary t with $a_2 < t < b_2$ and integrating $h(x_2)$ over the interval $[a_2, t]$, we obtain

$$\int_{a_2}^{t} h(x_2)\,dx_2 = \int_{a_2}^{t}\left[\int_{a_1}^{b_1} \frac{\partial f(x_1, x_2)}{\partial x_2} dx_1\right] dx_2. \tag{5.31}$$

The order of integration in (5.31) can be reversed by Fubini's theorem, so that, using the FTC (i) in (2.143),

$$\begin{aligned}
\int_{a_2}^{t} h(x_2)\,dx_2 &\overset{Fubini}{=} \int_{a_1}^{b_1}\left[\int_{a_2}^{t} \frac{\partial f(x_1, x_2)}{\partial x_2} dx_2\right] dx_1 \\
&\overset{FTC}{=} \int_{a_1}^{b_1} [f(x_1, t) - f(x_1, a_2)]\,dx_1 \\
&= \int_{a_1}^{b_1} f(x_1, t) dx_1 - \int_{a_1}^{b_1} f(x_1, a_2) dx_1.
\end{aligned} \tag{5.32}$$

From the FTC (ii, b) in (2.146) and differentiating both sides of (5.32) with respect to t, we obtain

$$\frac{d}{dt}\int_{a_2}^{t} h(x_2)\,dx_2 = h(t) = \int_{a_1}^{b_1} D_2 f(x_1, t)\,dx_1 \overset{(5.32)}{=} \frac{d}{dt}\int_{a_1}^{b_1} f(x_1, t)\,dx_1.$$

But the lhs of this is (5.30); the rhs is the lhs of (5.29); and the rhs of (5.29) is $h(t)$, thus showing the result. ■

Example 5.12 *Let $f : \mathbb{R}^2 \to \mathbb{R}, (x, y) \mapsto y^2 e^{2x}$. From (5.13), the fact that f is continuous implies that it is Riemann integrable on bounded rectangles. Let $a_1 = a_2 = 0$ and $b_1 = b_2 = 1$. If $D = [a_1, b_1] \times [a_2, b_2]$ we have*

$$\begin{aligned}
\int_D f &= \int_0^1 \int_0^1 y^2 e^{2x}\,dx\,dy = \int_0^1 \left[y^2 \frac{1}{2} e^{2x}\right]_{x=0}^{x=1} dy = \int_0^1 y^2 \frac{1}{2} e^2 - y^2 \frac{1}{2} 1\,dy \\
&= \left[\frac{1}{3} y^3 \left(\frac{1}{2} e^2 - \frac{1}{2}\right)\right]_{y=0}^{y=1} = \frac{1}{6}(e^2 - 1).
\end{aligned}$$

The same result can be easily derived when interchanging the order of integration. However, in this example, the calculations can be simplified by factorizing *the integrated function:*

$$\begin{aligned}
\int_D f &= \int_0^1 \int_0^1 y^2 e^{2x}\,dx\,dy = \int_0^1 y^2 \int_0^1 e^{2x}\,dx\,dy = \int_0^1 e^{2x}\,dx \int_0^1 y^2\,dy \\
&= \left[\frac{1}{2} e^{2x}\right]_0^1 \left[\frac{1}{3} y^3\right]_0^1 = \frac{1}{2}(e^2 - 1)\frac{1}{3}.
\end{aligned}$$

Usually, a factorization will not be possible. ■

We state now the generalization of (5.28) to higher dimensions.

Theorem (Fubini): Suppose that the function $f : \mathbf{I} \to \mathbb{R}$ is integrable, where $\mathbf{I} = \mathbf{I_x} \times \mathbf{I_y}$ is a generalized rectangle in $\mathbb{R}^{n+k}$. For each point $\mathbf{x}$ in $\mathbf{I_x}$, define the function $F_\mathbf{x} : \mathbf{I_y} \to \mathbb{R}$ by

$$F_\mathbf{x}(\mathbf{y}) = f(\mathbf{x}, \mathbf{y}) \quad \text{for } \mathbf{y} \text{ in } \mathbf{I_y};$$

suppose that the function $F_\mathbf{x} : \mathbf{I_y} \to \mathbb{R}$ is integrable, and define

$$A(\mathbf{x}) = \int_{\mathbf{I_y}} f(\mathbf{x}, \mathbf{y}) d\mathbf{y}.$$

Then the function $A : \mathbf{I_x} \to \mathbb{R}$ is integrable, and

$$\int_\mathbf{I} f = \int_{\mathbf{I_x}} A(\mathbf{x}) d\mathbf{x} = \int_{\mathbf{I_x}} \left[\int_{\mathbf{I_y}} f(\mathbf{x}, \mathbf{y}) d\mathbf{y} \right] d\mathbf{x}. \tag{5.33}$$

We are often interested in integration on more general regions, especially on sets that are unbounded, such as $\mathbb{R}^2$ or regions of the form $D_\infty := (-\infty, b_1] \times (-\infty, b_2]$. Recall the discussion of improper integrals in §2.4.3.

In order to define an integral on D_∞, let $D_1 \subset D_2 \subset \ldots$ be a sequence of closed and bounded rectangles with $\bigcup_{k \in \mathbb{N}} D_k = D_\infty$. Then we define

$$\int_{D_\infty} f := \lim_{k \to \infty} \int_{D_k} f,$$

whenever the rhs exists. Fubini's theorem still applies in these more general cases; in particular,

$$\int_{D_\infty} f = \int_{-\infty}^{b_2} \left[\int_{-\infty}^{b_1} f(x_1, x_2)\, dx_1 \right] dx_2 = \int_{-\infty}^{b_1} \left[\int_{-\infty}^{b_2} f(x_1, x_2)\, dx_2 \right] dx_1.$$

See, for example, Lang (1997, §13.3) for details.

Example 5.13 *As an example of a double integral for which the two corresponding iterated integrals are not equal, let* $f(x, y) = (2 - xy)xy \exp(-xy)$. *Then*

$$\int_0^1 \int_0^\infty f(x, y)\, dy\, dx = 0 \quad \textit{and} \quad \int_0^\infty \int_0^1 f(x, y)\, dx\, dy = 1.$$

This is from Cornfield (1969, p. 630), who refers to notes from Courant in 1936. ■

We now show another important extension of the baseline Fubini result.

Theorem: For continuous functions $h : [a, b] \to \mathbb{R}$ and $g : [a, b] \to \mathbb{R}$ with the property that $h(x) \leq g(x)$ for all points x in $[a, b]$, define

$$D = \{(x, y) \mid a \leq x \leq b,\ \ h(x) \leq y \leq g(x)\}.$$

Suppose that the function $f : D \to \mathbb{R}$ is continuous and bounded. Then

$$\int_D f = \int_a^b \left[\int_{h(x)}^{g(x)} f(x, y) dy \right] dx. \tag{5.34}$$

Proofs can be found in, e.g., Fitzpatrick, p. 500; and Terrell, p. 385.

The idea in (5.34) can be extended to any number of dimensions. For the triple integral case, we have from Terrell, p. 386:

Theorem: Let $\alpha(x)$ and $\beta(x)$ be continuous functions for $a \leq x \leq b$, with $\alpha(x) \leq \beta(x)$, and let $\gamma(x,y)$ and $\delta(x,y)$ be continuous functions for $a \leq x \leq b$, $\alpha(x) \leq y \leq \beta(x)$, with $\gamma(x,y) \leq \delta(x,y)$. Let

$$D = \left\{(x,y,z) \in \mathbf{R}^3 : a \leq x \leq b,\ \ \alpha(x) \leq y \leq \beta(x),\ \ \gamma(x,y) \leq z \leq \delta(x,y)\right\}.$$

If $f : D \to \mathbf{R}$ is continuous, then $\int_D f$ exists and

$$\int_D f = \int_a^b \left(\int_{\alpha(x)}^{\beta(x)} \left(\int_{\gamma(x,y)}^{\delta(x,y)} f(x,y,z)dz \right) dy \right) dx.$$

Example 5.14 *For $D = \{(x,y) \mid x^2 + y^2 \leq 1, y \geq 0\}$, define $f : D \to \mathbb{R}$ to be the constant function with value 1. Then*

$$D = \left\{(x,y) \mid -1 \leq x \leq 1, 0 \leq y \leq \sqrt{1-x^2}\right\},$$

and

$$\int_D f = \int_{-1}^{1} \left[\int_0^{\sqrt{1-x^2}} dy \right] dx = \int_{-1}^{1} \left[\sqrt{1-x^2}\right] dx = \frac{\pi}{2},$$

where this integral is resolved in Example 2.27. ■

Generalizing (5.33) and (5.34), we get (Fitzpatrick, p. 503):

Theorem: For a Jordan domain K in $\mathbb{R}^n$, let $h : K \to \mathbb{R}$ and $g : K \to \mathbb{R}$ be continuous bounded functions with the property that, $\forall \mathbf{x} \in K$, $h(\mathbf{x}) \leq g(\mathbf{x})$. Define

$$D = \left\{(\mathbf{x}, y) \text{ in } \mathbb{R}^{n+1} \mid \mathbf{x} \text{ in } K,\ \ h(\mathbf{x}) \leq y \leq g(\mathbf{x})\right\}.$$

Suppose that the function $f : D \to \mathbb{R}$ is continuous and bounded. Then

$$\int_D f = \int_K \left[\int_{h(\mathbf{x})}^{g(\mathbf{x})} f(\mathbf{x}, y)dy \right] d\mathbf{x}.$$

5.5 Leibniz' Rule

As above, let $a_1, a_2, b_1, b_2 \in \mathbb{R}$ with $a_1 < b_1$, $a_2 < b_2$, and let $D := [a_1, b_1] \times [a_2, b_2]$. Assume functions $f : D \to \mathbb{R}$ and $D_1 f$ are continuous. Also let λ and θ be differentiable functions defined on $[a_2, b_2]$ such that $\lambda(x), \theta(x) \in [a_1, b_1]$ for all $x \in [a_2, b_2]$ and define the function $A : [a_2, b_2] \to \mathbb{R}$ by

$$A(x) := \int_{\lambda(x)}^{\theta(x)} f(x, y)\, dy. \tag{5.35}$$

We wish to determine if this function is differentiable and, if so, derive an expression for its derivative. First, define the function $H : [a_1, b_1] \times [a_1, b_1] \times [a_2, b_2] \to \mathbb{R}$, depending on the three variables a, b and x, by

$$H(a, b, x) := \int_a^b f(x, y)\, dy.$$

Note that $A(x) = H(\lambda(x), \theta(x), x)$ for every $x \in [a_2, b_2]$. From the FTC (2.146), H is differentiable for any $a, b \in (a_2, b_2)$ and $x \in (a_1, b_1)$, with (also recall Example 2.20)

$$\begin{aligned} \frac{\partial H}{\partial b}(a, b, x) &= \frac{\partial}{\partial b} \int_a^b f(x, y)\, dy = f(x, b), \\ \frac{\partial H}{\partial a}(a, b, x) &= \frac{\partial}{\partial a} \int_a^b f(x, y)\, dy = -\frac{\partial}{\partial a} \int_b^a f(x, y)\, dy = -f(x, a), \end{aligned}$$

and from (5.27), as $D_1 f$ was assumed continuous,

$$\frac{\partial H}{\partial x}(a, b, x) = \frac{\partial}{\partial x} \int_a^b f(x, y)\, dy = \int_a^b \frac{\partial f(x, y)}{\partial x}\, dy.$$

From the chain rule (4.97), it follows that, for $a_1 < x < b_1$, A is differentiable and

$$\begin{aligned} A'(x) &= \frac{\partial H}{\partial \lambda}\frac{d\lambda}{dx} + \frac{\partial H}{\partial \theta}\frac{d\theta}{dx} + \frac{\partial H}{\partial x}\frac{dx}{dx} \\ &= -f\left(x, \lambda(x)\right) \lambda'(x) + f\left(x, \theta(x)\right) \theta'(x) + \int_{\lambda(x)}^{\theta(x)} \frac{\partial f(x, y)}{\partial x}\, dy. \end{aligned} \tag{5.36}$$

Formula (5.36) is sometimes called "Leibniz' rule for differentiating an integral".

Example 5.15 *Consider the function $f(t) := \int_0^t e^{st}\, ds$. There are two possible ways to calculate its derivative at $t = 1$. Firstly, let us integrate in step one and then differentiate afterwards. For $t > 0$,*

$$f(t) = \left[\frac{1}{t} e^{st}\right]_0^t = \frac{1}{t}\left(e^{t^2} - 1\right),$$

and

$$f'(t) = \frac{-1}{t^2}\left(e^{t^2} - 1\right) + \frac{1}{t} 2te^{t^2} = \frac{-1}{t^2}\left(e^{t^2} - 1\right) + 2e^{t^2},$$

so that $f'(1) = -(e - 1) + 2e = 1 + e$.

Secondly, we can differentiate first, using Leibniz' rule, and then integrate in a second step. For $t > 0$, with $d(e^{st})/dt = se^{st}$ continuous,

$$f'(t) = \int_0^t se^{st} ds + 1 \cdot e^{t^2},$$

hence

$$f'(1) = \int_0^1 se^s ds + e^1 = [(s-1)e^s]_0^1 + e = 0 - (-1) + e = 1 + e.$$

In this case, Leibniz' rule saves us a bit of work. ■

Example 5.16 *(https://brilliant.org/wiki/differentiate-through-the-integral/). We wish to compute the definite integral*

$$\int_0^1 \frac{t^3 - 1}{\ln t} dt.$$

Recall from Example 2.12 that, for $f(x) = t^x$ and $t > 0$, $f'(x) = t^x = t^x \ln t$. Define

$$g(x) = \int_0^1 \frac{t^x - 1}{\ln t} dt,$$

and we wish to evaluate $g(3)$. Observe that the given integral has been recast as member of a family of definite integrals $g(x)$ indexed by the variable x. With

$$\frac{\partial}{\partial x} \frac{t^x - 1}{\ln t} = \frac{1}{\ln t} \frac{d}{dx} (t^x - 1) = t^x,$$

we have, by Leibniz' rule, or, in this case, just from (5.27),

$$g'(x) = \int_0^1 \frac{\partial}{\partial x} \frac{t^x - 1}{\ln t} dt = \int_0^1 t^x dt = \frac{t^{x+1}}{x+1}\Big|_0^1 = \frac{1}{x+1}.$$

Recalling Example 2.31, it follows that $g(x) = \ln|x+1| + C$ for some constant C. To determine C, note that $g(0) = 0$, so $0 = g(0) = \ln 1 + C = C$. Hence, $g(x) = \ln|x+1|$ for all x such that the integral exists. In particular, $g(3) = \ln 4 = 2 \ln 2$. ■

Although Leibniz' rule (5.36) follows directly from the multivariate chain rule, the expression itself does not appear to hold much intuition, and one might wonder how the formula could have been postulated without knowledge of the chain rule. It turns out that, with the right geometric representation, the formula is essentially obvious! This pleasant result is due to Frantz (2001), on which the following is based. Firstly, it suffices to set the lower limit $\lambda(x)$ in (5.35) to zero because, from (2.138),

$$A(x) = \int_{\lambda(x)}^{\theta(x)} f(x,y)\, dy = \int_0^{\theta(x)} f(x,y)\, dy - \int_0^{\lambda(x)} f(x,y)\, dy.$$

The first graphic[35] in Figure 37 shows a cross-section, or "slice" (or *lamina*) of A at a particular value of x, with $y = \theta(x)$ lying in the xy-plane. The second graphic also shows the lamina at $x + \triangle x$, with area $A(x + \triangle x)$, so that the change in height of the lamina, for any y, is $f(x + \triangle x, y) - f(x,y) \approx D_1 f(x,y) \triangle x =: f_x(x,y) \triangle x$. Similarly, the width of the lamina increases by approximately $\theta'(x) \triangle x$.

Figure 38 isolates this lamina for clarity, and defines regions A_1 and A_2. The change in the lamina's area, $\triangle A$, is then $A_1 + A_2$, plus the upper-right corner, which, compared to

[35] I am very grateful to Marc Frantz, the author of Frantz (2001), for constructing and providing me with the three graphs shown in the figures.

the size of A_1 and A_2, can be ignored (it becomes negligible much faster than A_1 and A_2 as $\triangle x \to 0$). Thus,

$$A_1 \approx \int_0^{\theta(x)} f_x(x,y)\,dy\,\triangle x \quad \text{and} \quad A_2 \approx f(x,\theta(x))\,\theta'(x)\,\triangle x,$$

i.e., dividing by $\triangle x$ gives

$$\frac{\triangle A}{\triangle x} \approx \int_0^{\theta(x)} f_x(x,y)\,dy + f(x,\theta(x))\theta'(x),$$

which is indeed $A'(x)$ in (5.36) with $\lambda\,(x) = 0$.

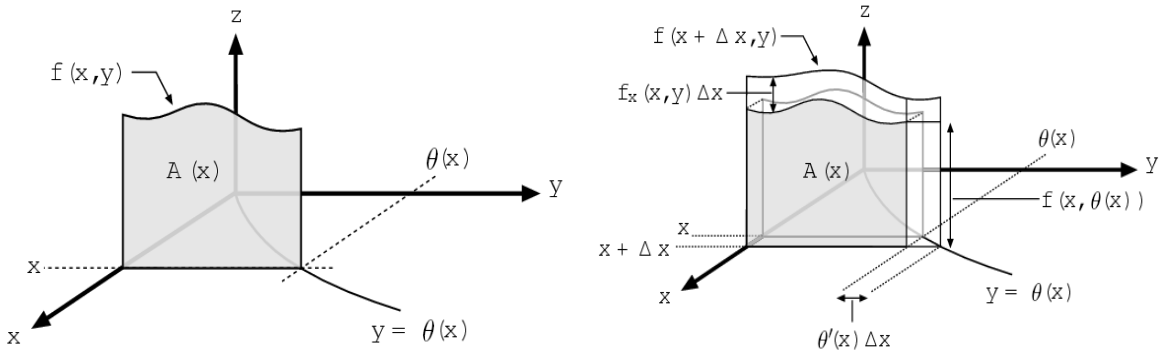


Figure 37: Geometric motivation for Leibniz' rule. Based on plots in M. Frantz, Visualizing Leibniz's Rule, *Mathematics Magazine*, 2001, 74(2):143–144.

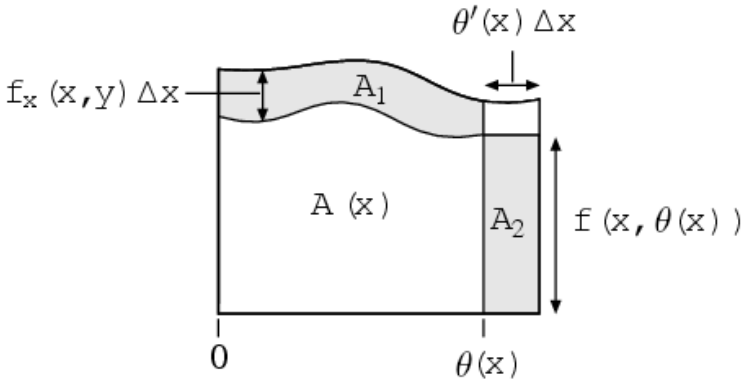


Figure 38: Magnified view of the relevant part of the second panel in Figure 37.

5.6 Integral Transformations, Polar and Spherical Coordinates

Theorem: Let $f : D \to \mathbb{R}$ be a continuous function, where domain D is an open subset of $\mathbb{R}^n$ with typical element $\mathbf{x} = (x_1, \ldots, x_n)$. Let $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), \ldots, g_n(\mathbf{x}) : D \to \mathbb{R}^n$ be a differentiable bijection with nonvanishing Jacobian

$$\mathbf{J} = \begin{pmatrix} \dfrac{\partial x_1}{\partial y_1} & \dfrac{\partial x_1}{\partial y_2} & \cdots & \dfrac{\partial x_1}{\partial y_n} \\ \dfrac{\partial x_2}{\partial y_1} & \dfrac{\partial x_2}{\partial y_2} & & \dfrac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial x_n}{\partial y_1} & \dfrac{\partial x_n}{\partial y_2} & \cdots & \dfrac{\partial x_n}{\partial y_n} \end{pmatrix}, \quad \text{where} \quad \mathbf{y} = \mathbf{g}(\mathbf{x}).$$

Then, for $S \subset D$,

$$\int_S f(\mathbf{x})\, d\mathbf{x} = \int_{\mathbf{g}(S)} f\left(\mathbf{g}^{-1}(\mathbf{y})\right) |\det \mathbf{J}|\, d\mathbf{y}, \tag{5.37}$$

where $\mathbf{g}(S) = \{\mathbf{y} : \mathbf{y} = \mathbf{g}(\mathbf{x}), \mathbf{x} \in S\}$.

This is referred to as the *multivariate change of variable formula*. The is a well-known and fundamental result in analysis, the rigorous proof of which, however, is somewhat involved.

Example 5.17 *This example is a special case of the next example, and is shown to help illustrate the idea in just two dimensions. Let* $I = \int_T \exp(x_1 + x_2)\, dx_1 dx_2$, *where, for* $a_1, b_1, a_2, b_2 \in \mathbb{R}$ *and* $T = \{a_1 \le x_1 \le b_1,\ a_2 \le x_1 + x_2 \le b_2\}$. *For* $\mathbf{X} = (x_1, x_2)$, *Let* $\mathbf{Y} = (y_1, y_2) = F(\mathbf{X})$, *where* $\mathbf{F} = (f_1(\mathbf{X}), f_2(\mathbf{X}))$, *with* $y_1 = f_1(\mathbf{X}) = x_1$ *and* $y_2 = f_2(\mathbf{X}) = x_1 + x_2$. *Denote the bijective inverse function as* $\mathbf{X} = G(\mathbf{Y}) = F^{-1}(\mathbf{Y}) = (g_1(\mathbf{Y}), g_2(\mathbf{Y}))$, *where* $g_1(\mathbf{Y}) = y_1$ *and* $g_2(\mathbf{Y}) = y_2 - y_1$. *Then*

$$\mathbf{J} = \begin{bmatrix} \partial x_1/\partial y_1 & \partial x_1/\partial y_2 \\ \partial x_2/\partial y_1 & \partial x_2/\partial y_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}, \quad \det(\mathbf{J}) = 1,$$

and the range of $\mathbf{Y}$ *is* $S = F(T) = [a_1, b_1] \times [a_2, b_2]$. *Thus,*

$$I = \int_S \exp(y_2)\, dy_1 dy_2 = \int_{a_1}^{b_1} dy_1 \int_{a_2}^{b_2} \exp(y_2)\, dy_2 = (b_1 - a_1)\left(e^{b_2} - e^{a_2}\right). \qquad \blacksquare$$

Example 5.18 *(Trench, 2013, Example 7.3.8) Evaluate*

$$I = \int_T e^{x_1 + x_2 + \cdots + x_n} d(x_1, x_2, \ldots, x_n),$$

where T *is the region defined by*

$$a_i \le x_1 + x_2 + \cdots + x_i \le b_i, \quad 1 \le i \le n.$$

Solution: We define the new variables $y_1, y_2, \ldots, y_n$ *by* $\mathbf{Y} = \mathbf{F}(\mathbf{X})$, *where*

$$f_i(\mathbf{X}) = x_1 + x_2 + \cdots + x_i, \quad 1 \le i \le n.$$

If $\mathbf{G}(\mathbf{Y}) = \mathbf{F}^{-1}(\mathbf{Y})$, *then* $T = \mathbf{G}(S)$, *where*

$$S = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n],$$

and $\mathbf{J} = 1$. *Hence,*

$$I = \int_S e^{y_n} d(y_1, y_2, \ldots, y_n) \tag{5.38}$$

$$= \int_{a_1}^{b_1} dy_1 \int_{a_2}^{b_2} dy_2 \cdots \int_{a_{n-1}}^{b_{n-1}} dy_{n-1} \int_{a_n}^{b_n} e^{y_n} dy_n \tag{5.39}$$

$$= (b_1 - a_1)(b_2 - a_2) \cdots (b_{n-1} - a_{n-1}) \left(e^{b_n} - e^{a_n}\right). \quad \blacksquare \tag{5.40}$$

Consider the special case of (5.37) using polar coordinates, i.e., $x = c_1(r, \theta) = r\cos\theta$ and $y = c_2(r, \theta) = r\sin\theta$. Then

$$\det \mathbf{J} = \begin{vmatrix} \partial x/\partial r & \partial x/\partial \theta \\ \partial y/\partial r & \partial y/\partial \theta \end{vmatrix} = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r\cos^2\theta + r\sin^2\theta = r, \tag{5.41}$$

which is positive, so that (5.37) implies

$$\iint f(x, y)\, dxdy = \iint f(r\cos\theta, r\sin\theta)\; r\, dr\, d\theta. \tag{5.42}$$

Example 5.19 *This is a particularly useful transformation when* $f(x, y)$ *depends only on the distance measure* $r^2 = x^2 + y^2$ *and the range of integration is a circle centered around* $(0, 0)$. *For example, if* $f(x, y) = (k + x^2 + y^2)^p$ *for constants* k *and* $p \geq 0$, *and* S *is such a circle with radius* a, *then (with* $t = r^2$, $r = t^{1/2}$, $dr = \frac{1}{2}t^{-1/2}dt$*),*

$$I = \iint_S f(x, y)\; dx\, dy = \int_0^{2\pi} \int_0^a \left(k + r^2\right)^p r\, dr\, d\theta = \frac{1}{2}\int_0^{2\pi} \int_0^{a^2} (k + t)^p\, t^{1/2} t^{-1/2}\, dt\, d\theta$$

$$= \frac{1}{2}\int_0^{2\pi} d\theta \cdot \int_0^{a^2} (k + t)^p\, dt = \pi \int_0^{a^2} (k + t)^p\, dt = \frac{\pi}{p + 1}\left(\left(k + a^2\right)^{p+1} - k^{p+1}\right),$$

having used the substitution $s = k + t$ *in the last equality. For* $k = p = 0$, I *reduces to* πa^2, *the area of a circle of radius* a. $\blacksquare$

The following example is very important, as it relates to the Gaussian distribution.

Example 5.20 *To compute* $I_1 = \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}t^2\right) dt = 2\int_0^{\infty} \exp\left(-\frac{1}{2}t^2\right) dt$, *let* $x^2 = t^2/2$ *so that*

$$I_1 = 2\sqrt{2}\int_0^{\infty} \exp\left(-x^2\right) dx = \sqrt{2}\int_{-\infty}^{\infty} \exp\left(-x^2\right) dx =: \sqrt{2} I_2.$$

Then, observe that

$$I_2^2 = \int_{-\infty}^{\infty} \exp\left(-x^2\right) dx \int_{-\infty}^{\infty} \exp\left(-y^2\right) dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)}\, dx\, dy,$$

or, transforming to polar coordinates,

$$I_2^2 = \int_0^{2\pi} \int_0^{\infty} \exp\left(-r^2\right) r\, dr\, d\theta = 2\pi \lim_{t\to\infty}\left(-\frac{1}{2}e^{-r^2}\Big|_0^t\right) = \pi,$$

so that $I_2 = \sqrt{\pi}$ *and* $I_1 = \sqrt{2\pi}$. $\blacksquare$

Example 5.21 *(https://brilliant.org/wiki/differentiate-through-the-integral/). We wish to compute* $\int_0^\infty e^{-x^2/2}dx$*; recall* (2.191)*. Define a function*

$$g(t) = \left(\int_0^t e^{-x^2/2}dx\right)^2.$$

Our goal is to compute $g(\infty)$ *and then take its square root. Differentiating with respect to* t *gives*

$$g'(t) = 2\cdot\left(\int_0^t e^{-x^2/2}dx\right)\cdot\left(\frac{d}{dt}\int_0^t e^{-x^2/2}dx\right) = 2e^{-t^2/2}\int_0^t e^{-x^2/2}dx = 2\int_0^t e^{-\left(t^2+x^2\right)/2}dx.$$

Make the change of variables $u = x/t$*, so that the integral transforms to*

$$g'(t) = 2\int_0^1 te^{-\left(1+u^2\right)t^2/2}du.$$

Now, the integrand has a closed-form antiderivative with respect to t *:*

$$g'(t) = -2\int_0^1 \frac{\partial}{\partial t}\frac{e^{-\left(1+u^2\right)t^2/2}}{1+u^2}du = -2\frac{d}{dt}\int_0^1 \frac{e^{-\left(1+u^2\right)t^2/2}}{1+u^2}du.$$

Set

$$h(t) = \int_0^1 \frac{e^{-\left(1+x^2\right)t^2/2}}{1+x^2}dx$$

Then by the above calculation, $g'(t) = -2h'(t)$*, so* $g(t) = -2h(t) + C$*. To determine* C*, take* $t \to 0$ *in the equation; since* $g(0) = 0$ *and*

$$h(0) = \int_0^1 \frac{1}{1+x^2}dx = \tan^{-1}x\Big|_0^1 = \frac{\pi}{4},$$

it follows that $0 = -\pi/2 + C \Longrightarrow C = \pi/2$*. Finally, taking* $t \to \infty$*, we conclude* $g(\infty) = -2h(\infty) + \pi/2 = \pi/2$*. Thus,* $\int_0^\infty e^{-x^2/2}dx = \sqrt{\frac{\pi}{2}}$*, which of course agrees with the result from Example 5.20.* ■

We now turn to a useful change of variables formula in three dimensions. For each point $\mathbf{u} = (x, y, z)$ in $\mathbb{R}^3$ that does not lie on the z-axis, we define $\rho = \sqrt{x^2+y^2+z^2}$. It is not difficult to see that there are unique numbers θ in the interval $[0, 2\pi)$ and ϕ in the interval $(0, \pi)$ such that

$$\mathbf{u} = (x, y, z) = (\rho\sin\phi\cos\theta, \rho\sin\phi\sin\theta, \rho\cos\phi).$$

The triple of numbers (ρ, ϕ, θ) is called a choice of spherical coordinates for the point $\mathbf{u}$. Define $\mathcal{O}$ to be the open subset of $\mathbb{R}^3$ consisting of points (ρ, ϕ, θ) with $\rho > 0$, $0 < \phi < \pi$, and $0 < \theta < 2\pi$ and then define $\Psi : \mathcal{O} \to \mathbb{R}^3$ by

$$\Psi(\rho, \phi, \theta) = (\rho\sin\phi\cos\theta, \rho\sin\phi\sin\theta, \rho\cos\phi) \quad \text{for } (\rho, \phi, \theta) \text{ in } \mathcal{O}.$$

See Figure 39. It is clear that the mapping $\Psi : \mathcal{O} \to \mathbb{R}^3$ is both continuously differentiable and one-to-one. Also, at each point (ρ, ϕ, θ) in $\mathcal{O}$, the derivative matrix is given by

$$\mathbf{D}\Psi(\rho, \phi, \theta) = \begin{pmatrix} \sin\phi\cos\theta & \rho\cos\phi\cos\theta & -\rho\sin\phi\sin\theta \\ \sin\phi\sin\theta & \rho\cos\phi\sin\theta & \rho\sin\phi\cos\theta \\ \cos\phi & -\rho\sin\phi & 0 \end{pmatrix}.$$

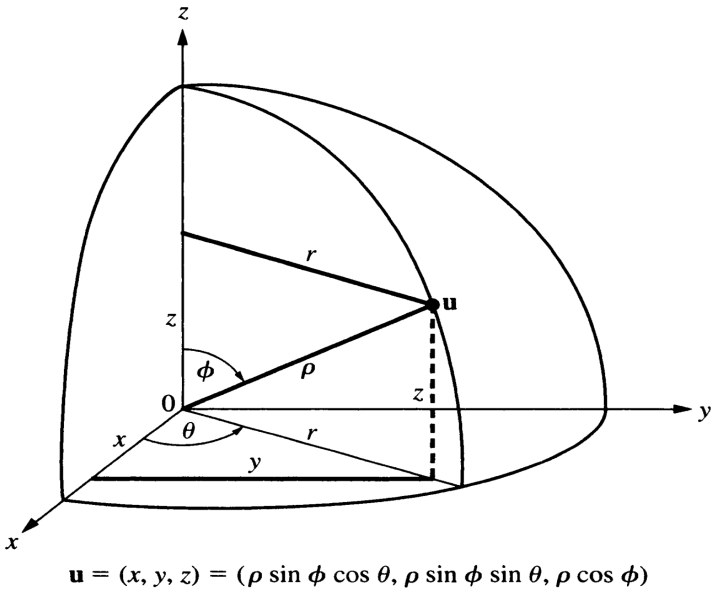


Figure 39: Spherical coordinates. From Fitzpatrick, p. 508.

A brief computation yields $\det \mathbf{D}\Psi(\rho,\phi,\theta) = \rho^2 \sin\phi \neq 0$. Thus, the derivative matrix $\mathbf{D}\Psi(\rho,\phi,\theta)$ is invertible, so $\Psi : \mathcal{O} \to \mathbb{R}^3$ is a smooth change of variables. For $0 < \rho_1 < \rho_2$, $0 < \phi_1 < \phi_2 < \pi$, and $0 < \theta_1 < \theta_2 < 2\pi$, define $K = [\rho_1, \rho_2] \times [\phi_1, \phi_2] \times [\theta_1, \theta_2]$.

Suppose that the function $f : \Psi(K) \to \mathbb{R}$ is continuous. Then by the integral transformation formula (5.37) and Fubini's Theorem,

$$\begin{aligned}
&\int_{\Psi(K)} f(x,y,z)\,dx\,dy\,dz \\
&\quad = \int_K \left[f(\rho\sin\phi\cos\theta, \rho\sin\phi\sin\theta, \rho\cos\phi)\rho^2\sin\phi \right]\, d\rho\,d\phi\,d\theta \\
&\quad = \int_{\theta_1}^{\theta_2} \left[\int_{\phi_1}^{\phi_2} \left\{ \int_{\rho_1}^{\rho_2} f(\rho\sin\phi\cos\theta, \rho\sin\phi\sin\theta, \rho\cos\phi)\rho^2\sin\phi\, d\rho \right\} d\phi \right] d\theta. \qquad (5.43)
\end{aligned}$$

Example 5.22 *For $a > 0$, we find the volume of the ball in $\mathbb{R}^3$ of radius a,*

$$B_a = \left\{ (x,y,z) \mid x^2 + y^2 + z^2 \leq a^2 \right\}.$$

Indeed, by formula (5.43),

$$\begin{aligned}
\operatorname{vol} B_a &= \int_{B_a} 1\,dx\,dy\,dz \\
&= \int_0^{2\pi} \left[\int_0^{\pi} \left\{ \int_0^{a} \rho^2\sin\phi\, d\rho \right\} d\phi \right] d\theta = [4/3]\pi a^3.
\end{aligned}$$

As Fitzpatrick, p. 509 states, Archimedes discovered the formula for the volume of a ball. He was so proud of this accomplishment that he had the formula inscribed on his tomb. ■

5.7 Multivariate Transformations for Random Variables

The Jacobian transformation is the key result required to compute the distribution of a set of random variables that are suitable functions of another set of random variables. The result is as follows:

Theorem: Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an n-dimensional continuous random variable and let function $\mathbf{g} = (g_1(\mathbf{x}), \ldots, g_n(\mathbf{x}))$ be a continuous bijection that maps $\mathcal{S}_\mathbf{X} \subset \mathbb{R}^n$, the support of $\mathbf{X}$, onto $\mathcal{S}_\mathbf{Y} \subset \mathbb{R}^n$. Then the probability density function (pdf) of $\mathbf{Y} = (Y_1, \ldots, Y_n) = \mathbf{g}(\mathbf{X})$ is given by

$$f_\mathbf{Y}(\mathbf{y}) = f_\mathbf{X}(\mathbf{x}) \left|\det \mathbf{J}\right|, \tag{5.44}$$

where $\mathbf{x} = \mathbf{g}^{-1}(\mathbf{y}) = \left(g_1^{-1}(\mathbf{y}), \ldots, g_n^{-1}(\mathbf{y})\right)$ and

$$\mathbf{J} = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & & \frac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{pmatrix} = \begin{pmatrix} \frac{\partial g_1^{-1}(\mathbf{y})}{\partial y_1} & \frac{\partial g_1^{-1}(\mathbf{y})}{\partial y_2} & \cdots & \frac{\partial g_1^{-1}(\mathbf{y})}{\partial y_n} \\ \frac{\partial g_2^{-1}(\mathbf{y})}{\partial y_1} & \frac{\partial g_2^{-1}(\mathbf{y})}{\partial y_2} & & \frac{\partial g_2^{-1}(\mathbf{y})}{\partial y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_n^{-1}(\mathbf{y})}{\partial y_1} & \frac{\partial g_n^{-1}(\mathbf{y})}{\partial y_2} & \cdots & \frac{\partial g_n^{-1}(\mathbf{y})}{\partial y_n} \end{pmatrix} \tag{5.45}$$

is the Jacobian of $\mathbf{g}$.

Notice that (5.44) reduces to the simple equation relevant for the univariate case.

Outline of proof: Let $h : \mathbb{R}^n \to \mathbb{R}$ be a bounded, measurable function so that

$$\mathbb{E}[h(\mathbf{Y})] = \mathbb{E}[h(\mathbf{g}(\mathbf{X}))] = \int_{\mathcal{S}_\mathbf{X}} h(\mathbf{g}(\mathbf{x})) f_\mathbf{X}(\mathbf{x}) \mathrm{d}\mathbf{x} = \int_{\mathcal{S}_\mathbf{Y}} h(\mathbf{y}) f_\mathbf{X}\left(\mathbf{g}^{-1}(\mathbf{y})\right) \left|\det \mathbf{J}\right| d\mathbf{y}.$$

In particular, let $h = \mathbb{I}_B(\mathbf{y})$ for a Borel set $B \in \mathcal{B}^n$, so that

$$\mathbb{E}[h(\mathbf{Y})] = \Pr(\mathbf{Y} \in B) = \int_{\mathcal{S}_\mathbf{Y}} \mathbb{I}_B(\mathbf{y}) f_\mathbf{X}\left(\mathbf{g}^{-1}(\mathbf{y})\right) \left|\det \mathbf{J}\right| d\mathbf{y}.$$

As this holds for all $B \in \mathcal{B}^n$, $f_\mathbf{X}\left(\mathbf{g}^{-1}(\mathbf{y})\right) \left|\det \mathbf{J}\right|$ is a probability density function for $\mathbf{Y}$.

The following examples are from Paolella, Fundamental Probability, Ch. 9).

Example 5.23 *Consider calculating the joint distribution of* $S = X + Y$ *and* $D = X - Y$ *and their marginals for* $X, Y \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$. *Adding and subtracting the two equations yields* $X = (S + D)/2$ *and* $Y = (S - D)/2$. *From these,*

$$\mathbf{J} = \begin{bmatrix} \partial x/\partial s & \partial x/\partial d \\ \partial y/\partial s & \partial y/\partial d \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{bmatrix}, \quad \det \mathbf{J} = -\frac{1}{2},$$

so that

$$f_{S,D}(s,d) = \left|\det \mathbf{J}\right| f_{X,Y}(x,y) = \frac{\lambda^2}{2} \exp\{-\lambda s\} \mathbb{I}_{(0,\infty)}(s+d) \mathbb{I}_{(0,\infty)}(s-d).$$

The constraints imply $s > d$ and $s > -d$; if $d < 0$ ($d > 0$) then $s > -d$ ($s > d$) is relevant. Thus,

$$\begin{aligned} f_D(d) &= \mathbb{I}_{(-\infty,0)}(d) \int_{-d}^{\infty} f_{S,D}(s,d)\, ds + \mathbb{I}_{(0,\infty)}(d) \int_{d}^{\infty} f_{S,D}(s,d)\, ds \\ &= \mathbb{I}_{(-\infty,0)}(d)\, \frac{\lambda}{2} e^{\lambda d} + \mathbb{I}_{(0,\infty)}(d)\, \frac{\lambda}{2} e^{-\lambda d} = \frac{\lambda}{2} e^{-\lambda |d|}, \end{aligned}$$

i.e., $D \sim \text{Lap}(0, \lambda)$. Next, $s + d > 0$ and $s - d > 0$ imply that $-s < d < s$, giving

$$f_S(s) = \int_{-s}^{s} \frac{\lambda^2}{2} \exp\{-\lambda s\}\, dd = \frac{\lambda^2}{2} \exp\{-\lambda s\} \int_{-s}^{s} dd = s\lambda^2 e^{-\lambda s} \mathbb{I}_{(0,\infty)}(s),$$

which follows because $S > 0$ from its definition. Thus, $S \sim \text{Gam}(2, \lambda)$, which agrees with the more general result regarding sums of iid exponentials. ■

Example 5.24 *Let $Z_i \overset{\text{iid}}{\sim} \text{N}(0,1)$, $i = 1, 2$. The inverse transformation and Jacobian of $S = Z_1 + Z_2$ and $D = Z_1 - Z_2$ was derived in Example 5.23, so that*

$$\begin{aligned} f_{S,D}(s,d) &= \frac{1}{2} f_{Z_1,Z_2}(z_1, z_2) \\ &= \frac{1}{2} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2} z_1^2\right\} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2} z_2^2\right\} \\ &= \frac{1}{2} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{s+d}{2}\right)^2\right\} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{s-d}{2}\right)^2\right\} \\ &= \frac{1}{2} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\frac{1}{2} d^2 + \frac{1}{2} s^2\right]\right\} \\ &= \frac{1}{\sqrt{2}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{d}{\sqrt{2}}\right)^2\right\} \frac{1}{\sqrt{2}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{s}{\sqrt{2}}\right)^2\right\}, \end{aligned}$$

i.e., S and D are independent, with $S \sim \text{N}(0, 2)$ and $D \sim \text{N}(0, 2)$. ■

Example 5.25 *Let $X_i \overset{i.i.d.}{\sim} \text{N}(0,1)$. Interest centers on the distribution of*

$$Y_1 = g_1(\mathbf{X}) = \sum_{i=1}^{n} X_i^2.$$

To derive it, let $Y_i = g_i(\mathbf{X}) = X_i, i = 2, 3, \ldots, n$, perform the multivariate transformation to get $f_{\mathbf{Y}}$, and then integrate out $Y_2, \ldots, Y_n$ to get the distribution of Y_1. The following steps can be used.

1. *First do the $n = 2$ case, for which the fact that*

$$\int \frac{1}{\sqrt{y_1 - y_2^2}} dy_2 = c + \arcsin\left(\frac{y_2}{\sqrt{y_1}}\right)$$

can be helpful.

Let $n = 2$, so that $X_2 = g_2^{-1}(\mathbf{Y}) = Y_2$ and $X_1 = g_1^{-1}(\mathbf{Y}) = \pm\sqrt{Y_1 - Y_2^2}$. Splitting this into two regions,

$$f_{\mathbf{Y}}(\mathbf{y}) = |\det \mathbf{J}_1| \, f_{\mathbf{X}}(x_1, x_2) \, \mathbb{I}_{(-\infty,0)}(x_1) + |\det \mathbf{J}_2| \, f_{\mathbf{X}}(x_1, x_2) \, \mathbb{I}_{(0,-\infty)}(x_1)$$

where

$$\mathbf{J}_1 = \begin{bmatrix} \partial x_1/\partial y_1 & \partial x_1/\partial y_2 \\ \partial x_2/\partial y_1 & \partial x_2/\partial y_2 \end{bmatrix} = \begin{bmatrix} -\left(y_1 - y_2^2\right)^{-1/2}/2 & y_2\left(y_1 - y_2^2\right)^{-1/2} \\ 0 & 1 \end{bmatrix}$$

and similarly for $\mathbf{J}_2$, and, in both cases, $|\mathbf{J}_i| = \left(y_1 - y_2^2\right)^{-1/2}/2$. Thus,

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{2}\left(y_1 - y_2^2\right)^{-1/2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(y_1 - y_2^2\right)} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y_2^2} + \text{ same}$$

or

$$f_{\mathbf{Y}}(\mathbf{y}) = \left(y_1 - y_2^2\right)^{-1/2} \frac{1}{2\pi} e^{-\frac{1}{2}y_1} \mathbb{I}_{\left(y_2^2,\infty\right)}(y_1),$$

where the indicator function follows from

$$Y_1 = X_1^2 + X_2^2 \quad \text{and} \quad 0 \le X_2^2 \le X_1^2 + X_2^2$$

or $0 \le Y_2^2 \le Y_1$. This also implies $-\sqrt{Y_1} \le Y_2 \le \sqrt{Y_1}$, from which we have

$$f_{Y_1}(y_1) = \int_{-\sqrt{y_1}}^{\sqrt{y_1}} f_{\mathbf{Y}}(\mathbf{y}) dy_2 = \frac{1}{2\pi} e^{-\frac{1}{2}y_1} \int_{-\sqrt{y_1}}^{\sqrt{y_1}} \left(y_1 - y_2^2\right)^{-1/2} dy_2.$$

From the hint,

$$\int \frac{1}{\sqrt{y_1 - y_2^2}} dy_2 = c + \arcsin\left(\frac{y_2}{\sqrt{y_1}}\right)$$

and, clearly, $\arcsin(1) = \pi/2$ *and* $\arcsin(-1) = -\pi/2$, *so that*

$$f_{Y_1}(y_1) = \frac{1}{2} \exp\left(-y_1/2\right) \mathbb{I}_{(0,\infty)}(y_1)$$

where the indicator follows from the definition of $Y_1 = \sum_{i=1}^2 X_i^2$. Thus, $Y_1 \sim \chi_2^2$.

2. *Simplify the following integral, which is used in the general case:*

$$J = \int_0^{y_0} u^{m/2-1} \left(y_0 - u\right)^{-1/2} du.$$

This integral is, using $v = (y_0 - u)/y_0, u = (1 - v)y_0, du = -y_0 dv$,

$$\begin{aligned} J &= \int_0^{y_0} u^{m/2-1} \left(y_0 - u\right)^{-1/2} du = -\int_1^0 \left((1-v)y_0\right)^{m/2-1} \left(vy_0\right)^{-1/2} y_0 dv \\ &= y_0^{(m-1)/2} \int_0^1 v^{(1/2)-1} (1-v)^{m/2-1} dv \\ &= y_0^{(m-1)/2} B\left(\frac{1}{2}, \frac{m}{2}\right) = y_0^{(m-1)/2} \frac{\Gamma(1/2)\Gamma(m/2)}{\Gamma((m+1)/2)}. \end{aligned}$$

3. For the general case, conduct the multivariate transformation to show that

$$f_{Y_1}(y_1) = \frac{1}{(2\pi)^{n/2}} \mathrm{e}^{-\frac{1}{2}y_1} \int \cdots \int_{\mathcal{S}} \left(y_1 - \sum_{i=2}^{n} y_i^2 \right)^{-\frac{1}{2}} dy_2 \cdots \mathrm{d}y_n$$

where

$$\mathcal{S} = \left\{ (y_2, \ldots, y_n) \in \mathbb{R}^{n-1} : 0 < \sum_{i=2}^{n} y_i^2 < y_1 \right\}.$$

Then use the following identity (due to Joseph Liouville (1809-1882) in 1839, which extended a result from Dirichlet):

> *Let $\mathcal{V}$ be a volume consisting of (i) $x_i \geq 0$ and (ii) $t_1 \leq \sum (x_i/a_i)^{p_i} \leq t_2$, and let f be a continuous function on (t_1, t_2). Then, with $r_i = b_i/p_i$ and $R = \sum r_i$,*
>
> $$\begin{aligned} &\int \cdots \int_{\mathcal{V}} x_1^{b_1-1} \cdots x_n^{b_n-1} f\left[\left(\frac{x_1}{a_1} \right)^{p_1} + \cdots + \left(\frac{x_n}{a_n} \right)^{p_n} \right] \mathrm{d}x_1 \cdots \mathrm{d}x_n \\ &= \frac{\prod a_i^{b_i} p_i^{-1} \Gamma(r_i)}{\Gamma(R)} \int_{t_1}^{t_2} u^{R-1} f(u) \mathrm{d}u \end{aligned} \tag{5.46}$$
>
> *For details on this and similar results, see Andrews, Askey and Roy (1999, Section 1.8) and Jones (2001, Chapter 9).*

From the result using $n = 2$, one might guess that the general case might lead to $Y_1 \sim \chi^2_n$, which is true. Now $X_i = g_i^{-1}(\mathbf{Y}) = Y_i, i = 2, \ldots, n$ and $X_1 = g_1^{-1}(\mathbf{Y}) = \pm\sqrt{Y_1 - \sum_{i=2}^{n} Y_i^2}$. We again need to split the support of $\mathbf{X}$ into two regions as before, but we have seen for the $n = 2$ case that the components are the same. Thus,

$$f_{\mathbf{Y}}(\mathbf{y}) = 2 \left| \mathbf{J}^{-1} \right|^{-1} f_{\mathbf{X}}(\mathbf{x}),$$

where we use $\mathbf{J}^{-1}$ instead of $\mathbf{J}$ because it is algebraically more convenient. We have

$$\mathbf{J}^{-1} = \begin{bmatrix} \partial y_1/\partial x_1 & \partial y_1/\partial x_2 & \cdots & \partial y_1/\partial x_n \\ \partial y_2/\partial x_1 & \partial y_2/\partial x_2 & \cdots & \partial y_2/\partial x_n \\ \vdots & \vdots & \ddots & \vdots \\ \partial y_n/\partial x_1 & \partial y_n/\partial x_2 & \cdots & \partial y_n/\partial x_n \end{bmatrix} = \begin{bmatrix} 2x_1 & 2x_2 & \cdots & 2x_n \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

with determinant $|\mathbf{J}^{-1}| = 2x_1 = 2\sqrt{Y_1 - \sum_{i=2}^{n} Y_i^2}$. Then, defining

$$D = y_1 - \sum_{i=2}^{n} y_i^2$$

for convenience, $|\mathbf{J}^{-1}|^{-1} = D^{-1/2}/2$ and

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= 2 \cdot \frac{1}{2} D^{-1/2} \frac{1}{(2\pi)^{n/2}} \exp\left\{ -\frac{1}{2} \left(D + \sum_{i=2}^{n} y_i^2 \right) \right\} \mathbb{I}_{\mathcal{S}_{\mathbf{Y}}}(\mathbf{y}) \\ &= D^{-1/2} \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}y_1} \mathbb{I}_{\mathcal{S}_{\mathbf{Y}}}(\mathbf{y}). \end{aligned} \tag{5.47}$$

(To check, with $n = 2$ and no indicator functions, this reduces to

$$f_{Y_1,Y_2}(y_1, y_2) = \left(y_1 - y_2^2\right)^{-1/2} \frac{1}{2\pi} e^{-\frac{1}{2}y_1},$$

which agrees with the direct derivation above.) Inserting D into (5.47) and setting up the integral,

$$f_{Y_1}(y_1) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}y_1} \int \cdots \int_{\mathcal{S}} \left(y_1 - \sum_{i=2}^{n} y_i^2\right)^{-\frac{1}{2}} dy_2 \cdots dy_n, \tag{5.48}$$

where

$$\mathcal{S} = \left\{(y_2, \ldots, y_n) \in \mathbb{R}^{n-1} : 0 < \sum_{i=2}^{n} y_i^2 < y_1\right\}.$$

We wish to apply Liouville's result to the integral

$$I = \int \cdots \int_{\mathcal{S}} \left(y_1 - \sum_{i=2}^{n} y_i^2\right)^{-\frac{1}{2}} dy_2 \cdots dy_n,$$

which we rewrite as

$$I = \int \cdots \int_{\mathcal{S}} \left(y_0 - \sum_{i=1}^{m} y_i^2\right)^{-\frac{1}{2}} dy_1 \cdots dy_m,$$

where

$$\mathcal{S} = \left\{(y_1, \ldots, y_m) \in \mathbb{R}^m : 0 < \sum_{i=1}^{m} y_i^2 < y_0\right\}$$

and $m = n - 1$. This is almost in the form of (9.13) when taking $p_i = 2$ and $a_i = b_i = 1, i = 1, \ldots, m$, (so that $r_i = 1/2$ and $R = m/2$) as well as $t_1 = 0, t_2 = y_0$ and $f(u) = (y_0 - u)^{-1/2}$.

The problem is that the condition $x_i \geq 0$ in (5.46) is not fulfilled. However, what we can compute is

$$I' = \int \cdots \int_{\mathcal{S}'} \left(y_0 - \sum_{i=1}^{m} y_i^2\right)^{-\frac{1}{2}} dy_1 \cdots dy_m,$$

where

$$\mathcal{S}' = \left\{(y_1, \ldots, y_m) \in \mathbb{R}_+^m : 0 < \sum_{i=1}^{m} y_i^2 < y_0\right\},$$

i.e., each y_i is restricted to be positive. Then, via the symmetry of the standard normal distribution about zero and the fact that each y_i enters the function f as y_i^2, we see that $I = 2^m I'$. Now, using (5.46),

$$I = 2^m I' = 2^m \frac{(1/2)^m \pi^{m/2}}{\Gamma(m/2)} \int_0^{y_0} u^{m/2-1} (y_0 - u)^{-1/2} du = \frac{\pi^{m/2}}{\Gamma(m/2)} J,$$

where the integral J was shown above to be

$$J = y_0^{(m-1)/2} \frac{\Gamma(1/2)\Gamma(m/2)}{\Gamma((m+1)/2)}.$$

Then, recalling that $m = n-1$,

$$I = \frac{\pi^{m/2}}{\Gamma(m/2)} y_0^{(m-1)/2} \frac{\Gamma(1/2)\Gamma(m/2)}{\Gamma((m+1)/2)} = \frac{\pi^{n/2}}{\Gamma(n/2)} y_0^{n/2-1}.$$

Renaming y_0 *back to* y_1, (5.48) *gives*

$$f_{Y_1}(y_1) = \frac{e^{-y_1/2}}{(2\pi)^{n/2}} \frac{\pi^{n/2}}{\Gamma(n/2)} y_1^{n/2-1} = \frac{1}{2^{n/2}\Gamma(n/2)} e^{-y_1/2} y_1^{n/2-1},$$

which is the χ^2_n *density.* ■

Our last example shows how a symbolic computing package (Maple) can be used to assist in some tedious but rudimentary calculations to expedite obtaining the solution.

Example 5.26 *Let* $Y_i \overset{\text{ind}}{\sim} \chi^2(d_i)$, $i = 1, \ldots, n$, *and define*

$$S_k = \frac{\sum_{i=1}^k Y_i}{\sum_{i=1}^k d_i}, \quad k = 1, \ldots, n-1, \quad \text{and} \quad F_k = \frac{Y_k/d_k}{S_{k-1}}, \quad k = 2, \ldots, n.$$

First consider the marginal distribution of F_k. *Recall that a* $\chi^2(d)$ *distribution is just a gamma distribution with shape* $d/2$ *and (inverse) scale 1/2. Thus,* $\sum_{i=1}^k Y_i \sim \chi^2\left(\sum_{i=1}^k d_i\right)$. *Further, recall that an F distribution arises as ratio of independent* χ^2 *r.v.s each divided by their degrees of freedom, so that* $F_k \sim F\left(d_k, \sum_{i=1}^{k-1} d_i\right)$.

Next, and more challenging, for $n = 3$, *we wish to show that* C, F_2 *and* F_3 *are independent, where* $C = \sum_{i=1}^n Y_i$. *This is true for general* n, *i.e., random variables*

$$\begin{aligned} C &= Y_1 + \cdots Y_n\,, \quad F_2 = \frac{Y_2/d_2}{Y_1/d_1}, \quad F_3 = \frac{Y_3/d_3}{(Y_1+Y_2)/(d_1+d_2)}, \ldots, \\ F_n &= \frac{Y_n/d_n}{(Y_1+Y_2+\cdots+Y_{n-1})/(d_1+d_2+\cdots+d_{n-1})}, \end{aligned}$$

are independent.[36]

For $n = 3$, *with Maple's assistance, we find*

$$Y_1 = \frac{C(d_1+d_2)d_1}{T_2T_3}, \quad Y_2 = \frac{CF_2d_2(d_1+d_2)}{T_2T_3}, \quad Y_3 = \frac{CF_3d_3}{T_3},$$

where $T_2 = (F_2d_2 + d_1)$ *and* $T_3 = F_3d_3 + d_1 + d_2$, *and*

$$\begin{bmatrix} \frac{\partial Y_1}{\partial C} & \frac{\partial Y_1}{\partial F_2} & \frac{\partial Y_1}{\partial F_3} \\ \frac{\partial Y_2}{\partial C} & \frac{\partial Y_2}{\partial F_2} & \frac{\partial Y_2}{\partial F_3} \\ \frac{\partial Y_3}{\partial C} & \frac{\partial Y_3}{\partial F_2} & \frac{\partial Y_3}{\partial F_3} \end{bmatrix} = \begin{bmatrix} \frac{d_1(d_1+d_2)}{T_3T_2} & -\frac{d_1d_2(d_1+d_2)C}{T_3T_2^2} & -\frac{d_1d_3(d_1+d_2)C}{T_3^2T_2} \\ \frac{F_2d_2(d_1+d_2)}{T_3T_2} & \frac{d_1d_2(d_1+d_2)C}{T_3T_2^2} & -\frac{F_2d_2d_3(d_1+d_2)C}{T_3^2T_2} \\ \frac{d_3F_3}{T_3} & 0 & \frac{d_3(d_1+d_2)C}{T_3^2} \end{bmatrix},$$

[36] The result is mentioned, for example, in Hogg and Tanis (1963, p. 436), who state that "this result is, in essence, well known and its proof, which is a rather easy exercise, is omitted". They use it in the context of sequential, or iterative, testing of the equality of exponential distributions. It is also used by Phillips and McCabe (1983) in the context of testing for structural change in the linear regression model.

with

$$\det \mathbf{J} = \frac{d_1 d_2 d_3 \left(d_1+d_2\right)^2 C^2}{T_3^3 T_2^2}.$$

Thus, with $d = d_1 + d_2 + d_3$ *(and, for simplicity, using* C, F_2 *and* F_3 *as both the names of the r.v.s and their arguments in the pdf), the joint density* $f_{C,F_2,F_3}\left(C, F_2, F_3\right)$ *is given by*

$$\begin{aligned}
& |\det \mathbf{J}| \prod_{i=1}^{3} \frac{1}{2^{d_i/2}\Gamma\left(d_i/2\right)} y_i^{d_i/2-1} e^{-y_i/2} \\
= \; & \frac{d_1 d_2 d_3 \left(d_1+d_2\right)^2 C^2}{T_3^3 T_2^2} \frac{\left(\frac{C(d_1+d_2)d_1}{T_2 T_3}\right)^{d_1/2-1} \left(\frac{CF_2 d_2 (d_1+d_2)}{T_2 T_3}\right)^{d_2/2-1} \left(\frac{CF_3 d_3}{T_3}\right)^{d_3/2-1}}{2^{d/2}\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)\Gamma\left(\frac{d_3}{2}\right)} \\
& \times \exp\left\{-\frac{1}{2}\left[\frac{C\left(d_1+d_2\right)d_1}{T_2 T_3} + \frac{CF_2 d_2\left(d_1+d_2\right)}{T_2 T_3} + \frac{CF_3 d_3}{T_3}\right]\right\} \\
= \; & \Gamma\left(d/2\right) \frac{d_1 d_2 d_3 \left(d_1+d_2\right)^2}{T_3^3 T_2^2} \frac{\left(\frac{(d_1+d_2)d_1}{T_2 T_3}\right)^{d_1/2-1} \left(\frac{F_2 d_2 (d_1+d_2)}{T_2 T_3}\right)^{d_2/2-1} \left(\frac{F_3 d_3}{T_3}\right)^{d_3/2-1}}{\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)\Gamma\left(\frac{d_3}{2}\right)} \\
& \times \frac{1}{2^{d/2}\Gamma\left(d/2\right)} C^{d/2-1} e^{-C/2}. \qquad (5.49)
\end{aligned}$$

The pdf of C *has been separated from the joint density in (5.49), showing that* $C \sim \chi^2\left(d\right)$ *and that* C *is independent of* F_2 *and* F_3*. It remains to simplify the joint pdf of* F_2 *and* F_3*. In this case, we have been told that* F_2 *and* F_3 *are independent, with* $F_k \sim F\left(d_k, \sum_{i=1}^{k-1} d_i\right)$*, so that we wish to confirm that*

$$\begin{aligned}
& \frac{\Gamma\left(\frac{d_1+d_2+d_3}{2}\right)\frac{d_3}{d_1+d_2}}{\Gamma\left(\frac{d_3}{2}\right)\Gamma\left(\frac{d_1+d_2}{2}\right)} \frac{\left(\frac{d_3}{d_1+d_2}F_3\right)^{d_3/2-1}}{\left(1+\frac{d_3}{d_1+d_2}F_3\right)^{(d_1+d_2+d_3)/2}} \times \frac{\Gamma\left(\frac{d_1+d_2}{2}\right)\left(\frac{d_2}{d_1}\right)}{\Gamma\left(\frac{d_2}{2}\right)\Gamma\left(\frac{d_1}{2}\right)} \frac{\left(\frac{d_2}{d_1}F_2\right)^{d_2/2-1}}{\left(1+\frac{d_2}{d_1}F_2\right)^{(d_1+d_2)/2}} \\
\stackrel{?}{=} \; & \Gamma\left(d/2\right) \frac{d_1 d_2 d_3 \left(d_1+d_2\right)^2}{T_3^3 T_2^2} \frac{\left(\frac{(d_1+d_2)d_1}{T_2 T_3}\right)^{d_1/2-1} \left(\frac{F_2 d_2 (d_1+d_2)}{T_2 T_3}\right)^{d_2/2-1} \left(\frac{F_3 d_3}{T_3}\right)^{d_3/2-1}}{\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)\Gamma\left(\frac{d_3}{2}\right)}.
\end{aligned}$$

The reader should check that, as $d = d_1 + d_2 + d_3$*, all the gamma terms can be cancelled from both sides. Verifying the equality of the remaining equation just entails simple algebra, and Maple indeed confirms that both sides are equal.* ■

6 Appendices

6.1 Further Material on the Gamma Function

Recall the two definitions of the gamma function $\Gamma(x)$ given in (1.38) and (1.42), which we repeat here:

$$\Gamma(x) := \int_0^\infty t^{x-1} e^{-t} dt, \quad x \in \mathbb{R}_{>0}, \tag{6.1}$$

and the Gauss product formula

$$\Gamma(x) = \lim_{n\to\infty} \frac{n! n^x}{x(x+1)\cdots(x+n)}, \quad x > 0. \tag{6.2}$$

We wish to prove their equivalence. We will require the beta function (1.49), i.e.,

$$B(a,b) := \int_0^1 x^{a-1} (1-x)^{b-1} dx, \quad a, b \in \mathbb{R}_{>0}, \tag{6.3}$$

and relationship (1.51), namely

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{6.4}$$

We will also require that $e^\lambda = \lim_{n\to\infty}(1+\lambda/n)^n$ from (2.104), and the result, shown in Example 2.59, that sequence $s_n = (1+1/n)^n$ is monotone increasing and bounded (which implies, it converges). The last tool we need is Dini's theorem (2.247) for sequences of functions $f_n : D \to \mathbb{R}$: If (i) the $f_n : D \to \mathbb{R}$ are continuous, (ii) the f_n are monotone, (iii) D is a closed, bounded interval, and (iv) $f_n \to f$ to $f \in \mathcal{C}^0$, then $f_n \rightrightarrows f$ on D.

Theorem: Integral expression (6.1) and product formula (6.2) are equivalent.

Proof: (Duren, Invitation to Classical Analysis, 2012, p. 255.) Observe that

$$\int_0^1 t^{x-1}(1-t)^n dt = B(x, n+1) = \frac{\Gamma(x)\Gamma(n+1)}{\Gamma(x+n+1)}$$
$$= \frac{n!\Gamma(x)}{(x+n)(x+n-1)\cdots x\Gamma(x)} = \frac{n!}{x(x+1)\ldots(x+n)}.$$

In the above integral, let $s = tn$, $t = s/n$, and $dt = (1/n)\,ds$, so that

$$\frac{n!}{x(x+1)\cdots(x+n)} = \int_0^1 t^{x-1}(1-t)^n dt = \frac{1}{n}\int_0^n \left(\frac{s}{n}\right)^{x-1}\left(1-\frac{s}{n}\right)^n ds$$
$$= \frac{1}{n^x}\int_0^n s^{x-1}\left(1-\frac{s}{n}\right)^n ds,$$

or, just replacing s with t,

$$\frac{n! n^x}{x(x+1)\cdots(x+n)} = \int_0^n t^{x-1}\left(1-\frac{t}{n}\right)^n dt = \int_0^\infty g_n(t) t^{x-1} dt,$$

where

$$g_n(t) = \begin{cases} \left(1-\frac{t}{n}\right)^n, & \text{for } 0 \le t \le n \\ 0, & \text{for } n < t < \infty. \end{cases}$$

Because $g_n(t)$ increases to the limit e^{-t} as $n \to \infty$, Dini's theorem ensures that the integrals converge and

$$\lim_{n\to\infty} \frac{n!n^x}{x(x+1)\cdots(x+n)} = \int_0^\infty e^{-t}t^{x-1}dt = \Gamma(x).$$

Recall that the gamma function provides a generalisation of the factorial function, i.e., from (1.40), for $n \in \mathbb{N}$, $\Gamma(n+1) = n!$. We wish to show that, if some function f satisfies the functional equation $f(1) = 1$ and $f(x+1) = xf(x)$ for $x > 0$, then in fact $f(x) = \Gamma(x)$, for $x > 0$. We will also show some related results for the beta function.

This material comes from Binmore, Mathematical Analysis: A Straightforward Approach, 2nd ed., 1982, predominantly §17.4-17.8. The desired result is shown below, after having proven some preliminary results.

Proposition (Binmore #12.21(6)): Let f be continuous on an interval I and satisfy

$$f\left(\frac{x+y}{2}\right) \leqslant \frac{f(x)+f(y)}{2}, \quad \forall x, y \in I.$$

Prove that, for any $x_1, x_2, \ldots, x_n$ in the interval I,

$$f\left(\frac{x_1+x_2+\ldots+x_n}{n}\right) \leqslant \frac{1}{n}\left\{f\left(x_1\right)+f\left(x_2\right)+\ldots+f\left(x_n\right)\right\}. \tag{6.5}$$

Proof: Let $P(n)$ be the assertion that

$$f\left(\frac{x_1+x_2+\ldots+x_n}{n}\right) \leqslant \frac{1}{n}\left\{f\left(x_1\right)+\cdots+f\left(x_n\right)\right\} \tag{6.6}$$

for any $x_1, x_2, \ldots, x_n$ in the interval I. That $P(2)$ holds is given. We assume that $P\left(2^n\right)$ holds and seek to establish $P\left(2^{n+1}\right)$. Write $m = 2^n$. Then $2m = 2^{n+1}$ and

$$\begin{aligned}
f\left(\frac{x_1+\ldots+x_{2m}}{2m}\right) &= f\left(\frac{1}{2}\left(\frac{1}{m}\left(x_1+\ldots+x_m\right)+\frac{1}{m}\left(x_{m+1}+\ldots+x_{2m}\right)\right\}\right) \\
&\leqslant \frac{1}{2}f\left(\frac{x_1+\ldots+x_m}{m}\right)+\frac{1}{2}f\left(\frac{x_{m+1}+\ldots+x_{2m}}{m}\right) \\
&\leqslant \frac{1}{2m}\left\{f\left(x_1\right)+\ldots+f\left(x_m\right)\right\}+\frac{1}{2m}\left\{f\left(x_{m+1}\right)+\ldots+f\left(x_{2m}\right)\right\} \\
&= \frac{1}{2m}\left\{f\left(x_1\right)+\ldots+f\left(x_{2m}\right)\right\}.
\end{aligned}$$

We now assume that $P(n)$ holds and seek to deduce that $P(n-1)$ holds. Write

$$X = \frac{x_1+x_2+\ldots+x_{n-1}}{n-1}$$

Then

$$\begin{aligned}
f\left(\frac{x_1+x_2+\ldots+x_{n-1}}{n-1}\right) = f(X) &= f\left(\frac{(n-1)X+X}{n}\right) \\
&= f\left(\frac{x_1+x_2+\ldots+x_{n-1}+X}{n}\right) \\
&\leqslant \frac{1}{n}\left\{f\left(x_1\right)+\ldots+f\left(x_{n-1}\right)+f(X)\right\}.
\end{aligned}$$

Hence

$$\left(1-\frac{1}{n}\right) f(X) \leqslant \frac{1}{n}\left\{f\left(x_1\right)+f\left(x_2\right)+\ldots+f\left(x_{n-1}\right)\right\}$$

and $P(n-1)$ follows.

Let α be a rational number satisfying $0<\alpha<1$. We may write $\alpha=m/n$. If $\beta=1-\alpha$, then $\beta=(n-m)/n$. Apply inequality (6.6) with $x_1=x_2=\ldots=x_m=x$ and $x_{m+1}=x_{m+2}=\ldots=x_n=y$. Then

$$f(\alpha x+\beta y) \leqslant \alpha f(x)+\beta f(y). \tag{6.7}$$

If α is irrational, consider a sequence $\langle\alpha_n\rangle$ of rational numbers such that $\alpha_n \to \alpha$ as $n \to \infty$. Then $\beta_n=1-\alpha_n \to 1-\alpha=\beta$ as $n \to \infty$. We have

$$f\left(\alpha_n x+\beta_n y\right) \leqslant \alpha_n f(x)+\beta_n f(y) \quad (n=1,2,\ldots).$$

Since f is continuous, consideration of the limit shows that (6.7) holds even when α is irrational.

Proposition (Binmore, §17.5, p. 159, #4): Show that $\Gamma(x+1)=x\Gamma(x)(x>0)$. Deduce that, for $n=1,2,3,\ldots$,

$$\Gamma(n+1)=n!. \tag{6.8}$$

Proof: Consider, for $0<\delta<\Delta$, the value of

$$\begin{aligned}\int_\delta^\Delta t^{x-1}e^{-t}dt &= \left[\frac{t^x}{x}e^{-t}\right]_\delta^\Delta+\int_\delta^\Delta \frac{t^x}{x}e^{-t}dt \\ &= \frac{1}{x}\left\{\Delta^x e^{-\Delta}-\delta^x e^{-\delta}\right\}+\frac{1}{x}\int_\delta^\Delta t^x e^{-t}dt.\end{aligned}$$

Observe that $\Delta^x e^{-\Delta} \to 0$ as $\Delta \to +\infty$ (exponentials drown powers) and $\delta^x e^{-\delta} \to 0$ as $\delta \to 0+$. It follows that

$$\Gamma(x)=\frac{1}{x}\Gamma(x+1),$$

as required.

Proposition (Binmore, §17.5, p. 159, #5): Prove that the gamma function is continuous on $(0,\infty)$. [Hint: If $0<\alpha<a \leq x \leqslant y \leq b<\beta$, prove that, for some constant H which does not depend on x or y, $|\Gamma(x)-\Gamma(y)| \leq H|x-y|\{\Gamma(\alpha)+\Gamma(\beta)\}$.]

Proof: Let $0<\alpha<a \leq x \leq y \leqslant b<\beta$ and let $0<\delta<\Delta$. Then

$$\left|\int_\delta^\Delta t^{x-1}e^{-t}dt-\int_\delta^\Delta t^{y-1}e^{-t}dt\right| \leq \int_\delta^\Delta \left|t^{x-1}-t^{y-1}\right| e^{-t}dt. \tag{6.9}$$

The Mean Value Theorem implies

$$\exists \xi \in (x,y) \quad \text{such that} \quad \frac{t^{x-1}-t^{y-1}}{x-y}=(\log t)t^{\xi-1}. \tag{6.10}$$

For any $r>0, t^{-r}\log t \to 0$ as $t \to +\infty$ and $t^r \log t \to 0$ as $t \to 0+$. It follows from (6.10) that we can find an H such that

$$\left|t^{x-1}-t^{y-1}\right| \leqslant H\left\{t^{\alpha-1}+t^{\beta-1}\right\}|x-y|$$

and hence it follows from (6.9) that

$$|\Gamma(x) - \Gamma(y)| \leqslant H|x - y|\{\Gamma(\alpha) + \Gamma(\beta)\}$$

and the continuity of the gamma function then follows from the sandwich theorem.

Proposition (Binmore, §17.5, p. 159, #6): Prove that the logarithm of the gamma function is convex on $(0, \infty)$.

Proof: We show that $\log \Gamma\left(\frac{1}{2}x + \frac{1}{2}y\right) \leqslant \frac{1}{2}\log\Gamma(x) + \frac{1}{2}\log\Gamma(y)$ and appeal to (6.5). By the Schwarz (Cauchy-Schwarz, Bunyakovsky-Schwarz) inequality (2.137), if $0 < \delta < \Delta$, then

$$\begin{aligned}\left\{\int_\delta^\Delta t^{(x+y-2)/2}e^{-t}dt\right\}^2 &= \left\{\int_\delta^\Delta \left(t^{(x-1)/2}e^{-t/2}\right)\left(t^{(y-1)/2}e^{-t/2}\right)dt\right\}^2 \\ &\leqslant \left\{\int_\delta^\Delta t^{x-1}e^{-t}dt\right\}\left\{\int_\delta^\Delta t^{y-1}e^{-t}dt\right\}.\end{aligned}$$

Thus

$$\left\{\Gamma\left(\frac{x+y}{2}\right)\right\}^2 \leqslant \{\Gamma(x)\}\{\Gamma(y)\},$$

and the result follows.

We are now in a position to prove the uniqueness result of the gamma function stated at the beginning.

Proposition (Binmore, p. 160): Let f be positive and continuous on $(0, \infty)$ and let its logarithm be convex on $(0, \infty)$. If f satisfies the functional equation $f(1) = 1$ and $f(x+1) = xf(x)$ for $x > 0$, then $f(x) = \Gamma(x)$, for $x > 0$.

Proof: The proof consists of showing that, under the hypotheses of the theorem, for each $x > 0$,

$$f(x) = \lim_{n\to\infty} \frac{n^x n!}{x(x+1)\dots(x+n)}. \tag{6.11}$$

It follows from exercise 17.5(4, 5 and 6) that the gamma function satisfies the hypotheses of the theorem. Since a sequence can have at most one limit, we can therefore conclude from (6.11) that $f(x) = \Gamma(x)(x > 0)$.

The proof of (6.11) uses the convexity of $\log f$. Suppose that $s \leqslant t \leqslant s + 1$. Then we may write $t = \alpha s + \beta(s+1)$ where $\alpha \geqslant 0, \beta \geqslant 0$ and $\alpha + \beta = 1$. Now $t = (\alpha+\beta)s + \beta = s + \beta$ and so $\beta = t - s$. From the convexity of $\log f$, it follows that

$$\begin{aligned}\log f(t) &\leq \alpha \log f(s) + \beta \log f(s+1) \\ f(t) &\leq \{f(s)\}^\alpha \{f(s+1)\}^\beta \\ &= \{f(s)\}^\alpha \{sf(s)\}^\beta \\ &= s^\beta f(s) = s^{t-s} f(s).\end{aligned} \tag{6.12}$$

Since $s \leqslant t \leqslant s + 1$, we also have $t - 1 \leq s \leq t$. Making appropriate substitutions in (6.12), we obtain

$$f(s) \leqslant (t-1)^{s-t+1} f(t-1) = (t-1)^{s-t} f(t). \tag{6.13}$$

Combining (6.12) and (6.13) yields the inequality

$$(t-1)^{t-s}f(s) \leqslant f(t) \leqslant s^{t-s}f(s).$$

Now suppose that $0 \leqslant x < 1$ and that n is a natural number. We may take $s = n+1$ and $t = x+n+1$. Then

$$(x+n)^x f(n+1) \leqslant f(x+n+1) \leqslant (n+1)^x f(n+1). \tag{6.14}$$

From this inequality it follows that

$$(x+n)^x n! \leqslant (x+n)(x+n-1)\dots xf(x) \leqslant (n+1)^x n!$$

or

$$\left(1+\frac{x}{n}\right)^x \leqslant \frac{(x+n)(x+n-1)\dots xf(x)}{n^x n!} \leqslant \left(1+\frac{1}{n}\right)^x.$$

This completes the proof of the formula (6.11) in the case when $0 < x \leq 1$. The general case is easily deduced with the help of the functional equation $f(x+1) = xf(x)$.

We now present some further results related to the gamma function, as well as some for the beta function.

Proposition (Binmore #17.8(1)): If $x > 0$, prove that

$$\Gamma(x) = \int_0^1 \left\{\log\frac{1}{t}\right\}^{x-1} dt$$

Proof: Make the change of variable $-t = \log u$ in the integral

$$\int_\delta^\Delta t^{x-1}e^{-t}dt.$$

Proposition (Binmore #17.8(3)): Use L'Hôpital's rule to show that

$$\lim_{z\to 0}\left\{\frac{\log(1+z)-z}{z^2}\right\} = -\frac{1}{2}.$$

Proof: We have

$$\begin{aligned}\lim_{z\to 0}\left\{\frac{\log(1+z)-z}{z^2}\right\} &= \lim_{z\to 0}\left\{\frac{(1+z)^{-1}-1}{2z}\right\} \\ &= \lim_{z\to 0}\left\{\frac{-(1+z)^{-2}}{2}\right\} = -\frac{1}{2}.\end{aligned}$$

Proposition (Binmore #17.8(4)): For the beta function $B(x,y)$ in (1.49) given by

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt,$$

confirm that the improper integral exists provided that $x > 0$ and $y > 0$. Prove that, for a given fixed value of $y > 0$, $B(x,y)$ is a positive, continuous function of x on $(0,\infty)$ whose logarithm is convex on $(0,\infty)$.

Proof: We use the Comparison Test for Improper Integrals, (2.197), to check that the improper integral exists for $x > 0$ and $y > 0$. The inequalities

$$t^{x-1}(1-t)^{y-1} < t^{x-1} \quad (0 < t < 1)$$
$$t^{x-1}(1-t)^{y-1} < (1-t)^{y-1} \quad (0 < t < 1)$$

suffice for this purpose. The fact that $B(x, y)$ is a continuous function of x (and of y) is proved in the same manner as the above proposition (§17.5, #5). The fact that its logarithm is convex is proved in the same manner as the above proposition (§17.5, #6).

6.2 The Digamma and Polygamma Functions

This material comes mostly from the math appendix in Paolella, Fundamental Probability: A Computational Approach.

The digamma function is given by

$$\psi(s) := \frac{d}{ds} \ln \Gamma(s) = \frac{\Gamma'(s)}{\Gamma(s)} = \int_0^\infty \left[\frac{e^{-t}}{t} - \frac{e^{-st}}{1-e^{-t}} \right] dt, \tag{6.15}$$

with the latter, well-known integral representation proven in, e.g., Andrews, Askey and Roy (1999, p. 26). Higher-order derivatives are denoted as

$$\psi^{(n)}(s) = \frac{d^n}{ds^n}\psi(s) = \frac{d^{n+1}}{ds^{n+1}} \ln \Gamma(s) = (-1)^{n+1} \int_0^\infty \frac{t^n e^{-st}}{1-e^{-t}} dt, \quad n \in \mathbb{N},$$

also known as the polygamma function. Numeric methods exist for their evaluation; see Abramowitz and Stegun (1972, Section 6.3).

A result of general interest (and required, e.g., to compute the variance of a Gumbel random variable) is that

$$\psi'(1) = \int_0^\infty \frac{te^{-t}}{1-e^{-t}} dt = \frac{\pi^2}{6};$$

see, e.g., Andrews, Askey and Roy (1999, p. 51 and 55) for proof.

Example 6.1 *We wish to determine the expectation of* $\ln X$*, where* X *is a chi-squared random variable, i.e., compute* $\mathbb{E}[\ln X]$ *when* $X \sim \chi^2_v$*. We could try to directly integrate based on the definition of expectation, i.e.,*

$$\mathbb{E}[\ln X] = \frac{1}{2^{v/2}\Gamma(v/2)} \int_0^\infty (\ln x) x^{v/2-1} e^{-x/2} dx, \tag{6.16}$$

but this seems to lead nowhere. Note instead that the m.g.f. of $Z = \ln X$ *is*

$$\mathbb{M}_Z(t) = \mathbb{E}\left[e^{tZ}\right] = \mathbb{E}\left[X^t\right] = \frac{1}{2^{v/2}\Gamma(v/2)} \int_0^\infty x^{t+v/2-1} e^{-x/2} dx,$$

or, with $y = x/2$*,*

$$\mathbb{M}_Z(t) = \frac{2^{t+v/2-1+1}}{2^{v/2}\Gamma(v/2)} \int_0^\infty y^{t+v/2-1} e^{-y} dy = 2^t \frac{\Gamma(t+v/2)}{\Gamma(v/2)}.$$

Then, with $d2^t/dt = 2^t \ln 2$ *(see Example 2.12),*

$$\frac{d}{dt}\mathbb{M}_Z(t) = \frac{1}{\Gamma(v/2)}\left(2^t\Gamma'(t+v/2) + 2^t \ln 2\Gamma(t+v/2)\right)$$

and

$$\mathbb{E}[\ln X] = \frac{d}{dt}\mathbb{M}_Z(t)\bigg|_{t=0} = \frac{\Gamma'(v/2)}{\Gamma(v/2)} + \ln 2 = \psi(v/2) + \ln 2.$$

Having seen the answer, the integral (6.16) is easy; differentiating $\Gamma(v/2)$ *with respect to* $v/2$*, using (2.97), and setting* $y = 2x$*,*

$$\begin{aligned}
\Gamma'\left(\frac{v}{2}\right) &= \int_0^\infty \frac{d}{d(v/2)} x^{v/2-1}e^{-x}dx = \int_0^\infty x^{v/2-1}(\ln x)e^{-x}dx \\
&= \int_0^\infty \left(\frac{y}{2}\right)^{v/2-1}\left(\ln\frac{y}{2}\right)e^{-y/2}\frac{dy}{2} \\
&= \frac{1}{2^{v/2}}\int_0^\infty y^{v/2-1}(\ln y)e^{-y/2}dy - \frac{\ln 2}{2^{v/2}}\int_0^\infty y^{v/2-1}e^{-y/2}dy \\
&= \Gamma(v/2)\mathbb{E}[\ln X] - (\ln 2)\Gamma(v/2),
\end{aligned}$$

giving $\mathbb{E}[\ln X] = \Gamma'(v/2)/\Gamma(v/2) + \ln 2$. ■

Example 6.2 *Computation of the polygamma function is of course available in numerical and symbolic algebra computer packages such as Matlab and Maple. Still, it is of interest to know how one could compute them without the use of these optimized routines. We need only consider* $s \in [1,2]$*, because, for* s *outside* $[1,2]$*, the recursion (Abramowitz and Stegun, 1972, eqs 6.3.5 and 6.4.6)*

$$\psi^{(n)}(s+1) = \psi^{(n)}(s) + (-1)^n n! s^{-(n+1)} \tag{6.17}$$

can be used. We begin with the expansions

$$\psi(s) = -\gamma + (s-1)\sum_{k=1}^{\infty}\frac{1}{k(k+s-1)}, \quad s \neq 0, -1, -2, \ldots, \tag{6.18}$$

and, for $n \in \mathbb{N}$*,*

$$\psi^{(n)}(s) = (-1)^{n+1}n!\sum_{k=0}^{\infty}(s+k)^{-(n+1)}, \quad s \neq 0, -1, -2, \tag{6.19}$$

which can be found in, e.g., Abramowitz and Stegun (1972, eqs 6.3.16 and 6.4.10), where γ *is Euler's constant,*

We will truncate these infinite expansions and approximate the tail sum by its continuity corrected integral. For example, for (6.19),

$$\psi^{(n)}(s) \simeq (-1)^{n+1}n!\left[\sum_{k=0}^{N_n}(s+k)^{-n-1} + \int_{N_n+\frac{1}{2}}^{\infty}\frac{dt}{(s+t)^{n+1}}\right].$$

This yields, for $1 \leq s \leq 2$ *and* $n \geq 1$*,*

$$\psi^{(n)}(s) \approx (-1)^{n+1}\left[n!\sum_{k=0}^{N_n}(s+k)^{-n-1} + \frac{1}{n}\left(s+N_n+\frac{1}{2}\right)^{-n}\right]. \tag{6.20}$$

For (6.18), we get

$$\psi(s) \approx -\gamma + (s-1)\sum_{k=1}^{N_0} k^{-1}(k+s-1)^{-1} + \ln\left|\frac{N_0+s-0.5}{N_0+0.5}\right|. \tag{6.21}$$

These approximations (6.20) and (6.21) are derived as follows. For (6.19), the integral is

$$\int_{N_n+\frac{1}{2}}^{\infty} \frac{dt}{(s+t)^{n+1}} = \int_{s+N_n+\frac{1}{2}}^{\infty} u^{-n-1}du = -\frac{1}{n}u^{-n}\Big|_{s+N_n+\frac{1}{2}}^{\infty} = \frac{1}{n}\left(s+N_n+\frac{1}{2}\right)^{-n}.$$

Similarly, for (6.18), substituting $u = 1+(s-1)/t$ leads to

$$\int_{N_0+\frac{1}{2}}^{\infty} \frac{s-1}{t(t+s-1)}dt = \int_{N_0+\frac{1}{2}}^{\infty} \frac{(s-1)/t^2}{1+(s-1)/t}dt = \ln\left(1+\frac{s-1}{N_0+\frac{1}{2}}\right),$$

so that

$$\psi(s) \approx -\gamma + (s-1)\sum_{k=1}^{N_0} k^{-1}(k+s-1)^{-1} + \ln\left|\frac{N_0+s-0.5}{N_0+0.5}\right|.$$

This technique is of use in general for approximating certain functions based on truncation of infinite expansions. ■

Example 6.3 *We wish to relate the digamma function (6.15) to the harmonic numbers $H_n = \sum_{k=1}^{n} k^{-1}$, and show that $\psi(1) = -\gamma$, where $\gamma = \lim_{n\to\infty}(\sum_1^n 1/n - \log n)$ is the Euler-Mascheroni constant, discussed in Example 2.64.*

Beginning with the second task, as in https://math.stackexchange.com/questions/4524968, from (1.42),

$$\Gamma(1+x) = x\Gamma(x) = \lim_{n\to\infty}\frac{n!n^x}{(1+x)\cdots(n+x)} = \lim_{n\to\infty} n^x \prod_{k=1}^{n}\frac{k}{k+x}.$$

Thus, for $|x|<1$,

$$\begin{aligned}
\log\Gamma(1+x) &= \lim_{n\to\infty}\left[x\log n - \sum_{k=1}^{n}\log\left(1+\frac{x}{k}\right)\right] \\
&= \lim_{n\to\infty}\left[x\log n + \sum_{k=1}^{n}\sum_{j=1}^{\infty}\frac{1}{j}\left(-\frac{x}{k}\right)^j\right] \\
&= \lim_{n\to\infty}\left[x\left(\log n - \sum_{k=1}^{n}\frac{1}{k}\right) + \sum_{j=2}^{\infty}\frac{(-x)^j}{j}\sum_{k=1}^{n}\frac{1}{k^j}\right] \\
&= -\gamma x + \sum_{j=2}^{\infty}\frac{(-x)^j}{j}\zeta(j),
\end{aligned}$$

where the zeta function is given in Example 2.57; and the exchange of limit and infinite sum follows from Tannery's theorem (2.233). Taking the derivative of both sides and evaluating at $x=0$ gives the result $\psi(1) = -\gamma$.

Next, as $\Gamma(z+1) = z\Gamma(z)$, taking logs gives $\ln(\Gamma(z+1)) = \ln(z)+\ln(\Gamma(z))$. Differentiating both sides with respect to z gives

$$\psi(z+1) = \psi(z) + \frac{1}{z},$$

from which it follows that $\psi(n) = H_{n-1} - \gamma$, where $H_0 = 0$. ■

6.3 Banach's Matchbox Problem

This section is not as relevant as the material in the main text, but it is instructive and interesting. It details the well-known "Banach's Matchbox Problem" from probability theory. It also shows an application of Wallis' product.

As the story goes, the mathematician Stefan Banach (1892–1945) kept two match boxes, one in each pocket, each originally containing N matches. Whenever he wanted a match, he randomly chose between the boxes (with equal probability) and took one out. Upon discovering an empty box, what is the probability that the other box contains $K = k$ matches, $k = 0, 1, \ldots, N$? (See Feller, 1968, p. 166).

Assume that he discovers the right hand pocket to be empty (rhpe). Because trials were random, X, the number of matches that were drawn from the left, can be thought of as "failures" from a negative binomial-type experiment with $p = 1/2$ (and support only $\{0, 1, \ldots, N\}$ instead of $\{0 \cup \mathbb{N}\}$), where sampling continues until $r = N + 1$ "successes" (draws from the right pocket) occur. Thus $\Pr(K = k \cap \text{rhpe}) = \Pr(X = x \cap \text{rhpe})$, where $X = N - K$, $x = N - k$ and

$$\Pr(X = x \cap \text{rhpe}) = \binom{r + x - 1}{x} p^r (1 - p)^x = \binom{2N - k}{N} \left(\frac{1}{2}\right)^{2N+1-k}.$$

With $\Pr(X = x) = \Pr(X = x \cap \text{rhpe}) + \Pr(X = x \cap \text{lhpe})$ and from the symmetry of the problem,

$$f(k; N) = \Pr(K = k \mid N) = 2 \Pr(X = x \cap \text{rhpe}) = \binom{2N - k}{N} \left(\frac{1}{2}\right)^{2N-k}. \tag{6.22}$$

From (1.18) in Example 1.6, this mass function indeed sums to one.

The pmf (6.22) can also be expressed recursively as

$$\begin{aligned} \Pr(K = k) &= \binom{2N - k}{N} \left(\frac{1}{2}\right)^{2N-k} \\ &= 2 \frac{N - (k - 1)}{2N - (k - 1)} \binom{2N - (k - 1)}{N} \left(\frac{1}{2}\right)^{2N-(k-1)} \\ &= \frac{N - (k - 1)}{N - \frac{k-1}{2}} \Pr(K = k - 1), \end{aligned} \tag{6.23}$$

from which it is directly seen that $\Pr(K = 0) = \Pr(K = 1)$ and that $\Pr(K = k)$ decreases in k, $k \geq 1$. Recursion (6.23) also provides a way of calculating $\Pr(K = k)$ avoiding the computation of the gamma function.

One natural generalization of the original Banach matchbox problem is to allow for different numbers of matches in the left and right pockets, say N_1 and N_2, and probability p not necessarily $1/2$ of drawing from the left side. Derive the mass function $f(k; N_1, N_2, p)$ and construct a computer program, say `banach(n1,n2,p,sim)` that computes it, simulates the process `sim` times, and finally plots the true and simulated mass functions overlaid. As an example, the first panel in Figure 40 was produced with the Matlab code and function

```
vec=banach(30,10,1/2,10000);
text(3,0.07,'N_1=30, N_2=10, p=1/2','fontsize',14)
```

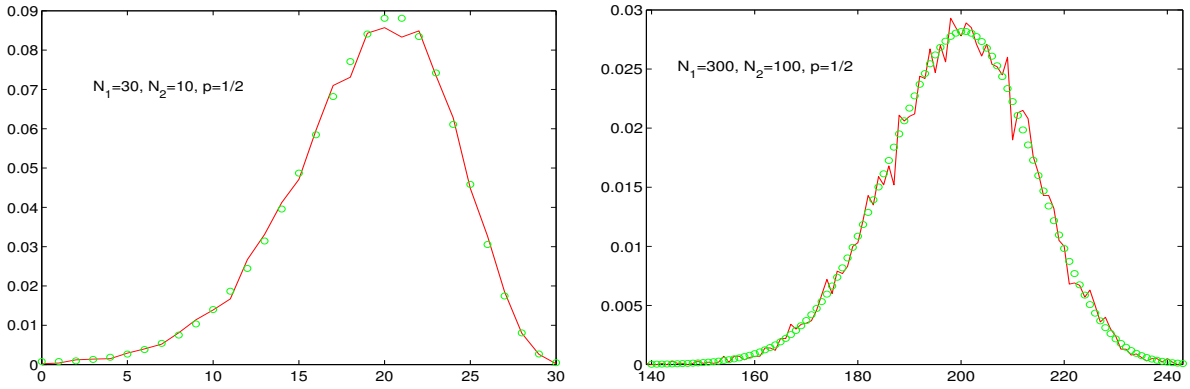


Figure 40: True (circles) and simulated (lines) mass function for the generalized Banach matchbox problem

For the mass function, let $X_i = N_i - K$, $x_i = N_i - k$, $i = 1, 2$, and $N = N_1 + N_2$, giving

$$\begin{aligned} f(k; N_1, N_2, p) &= \Pr(K = k \mid N_1, N_2, p) \\ &= \Pr(X_2 = x \cap \text{ lhpe }) + \Pr(X_1 = x \cap \text{ rhpe }) \\ &= \binom{N-k}{N_1} p^{N_1+1}(1-p)^{N_2-k} \mathbb{I}_{\{0,1,\ldots,N_2\}}(k) \\ &+ \binom{N-k}{N_2} (1-p)^{N_2+1} p^{N_1-k} \mathbb{I}_{\{0,1,\ldots,N_1\}}(k). \end{aligned}$$

Matlab code for the desired function is given in Listing 3. It imposes, without loss of generality, that $N_1 \geq N_2$.

Further extensions to the matchbox problem could allow for more than two matchboxes; see, for example, Cacoullos (1989, p. 80(317,318)) for some analytic results. Assume that Banach has $n \geq 2$ matchboxes distributed throughout his many pockets, each of which initially contains $N_i > 0$ matches, $i = 1, \ldots, n$. Associated with each box is a probability p_i, with $\sum_i p_i = 1$.

Listing 4 gives a Matlab program that simulates the process and, once an empty box is discovered, reports the minimum and maximum number of matches in the remaining matchboxes. This can be used to simulate the process, and, for example, plot (approximations to the exact) marginal mass functions, overlaid together on a single plot.

We wish to compute the expected value of the random variable K associated with Banach's matchbox problem. With $N = 1$, it is clear that $\mathbb{E}[K] = 1/2$. For general N, determining $\mathbb{E}_N[K]$ will exercise our combinatoric skills. Of use will be (1.16). The answer is

$$\mathbb{E}_N[K] = \binom{N + 1/2}{N} - 1, \tag{6.24}$$

as was found by Feller (1957, p. 212) using a different method of derivation than what we show here.

```
function vec=banach(n1,n2,p,sim)
vec=zeros(sim,1);
for i=1:sim
  vec(i)=simul(n1,n2,p); if mod(i,100)==0, i, end
end
tt=tabulate(vec+1); a=tt(:,2); mx=tt(end,1)-1; b=0:mx; a=a./sim;
true=echt(n1,n2,p); plot(b,a,'r-',0:mx,true(1:mx+1),'go')
mn=min(vec); ax=axis; axis([mn mx 0 ax(4)]), set(gca,'fontsize',14)

function echte=echt(n1,n2,p) % true mass function
mx=max(n1,n2); echte=zeros(1,mx); n=n1+n2; d=0:mx;
d1=0:n2; h1=c(n-d1,n1);
k1=p^(n1+1).*(1-p).^(n2-d1); h1=h1.*k1;
d2=0:n1; h2=c(n-d,n2);
k2=(1-p)^(n2+1).*p.^(n1-d2); h2=h2.*k2;
if n1>n2
  g=zeros(1,n1-n2); h1=[h1,g];
else
  g=zeros(1,n2-n1); h2=[h2,g];
end
echte=h1+h2;

function output=simul(n1,n2,p);
ok=1; zaehlerone=0; zaehlertwo=0;
while ok
  y=unifrnd(0,1,1,1);
  if y<p % box 1
    if n1==0, x=n2; ok=0; end
    if n1>0, n1=n1-1; zaehlerone=zaehlerone+1; end
  end
  if y>p
    if n2==0, x=n1; ok=0; end
    if n2>0, n2=n2-1; zaehlertwo=zaehlertwo+1; end
  end
end
output=x;
```

Program Listing 3: Code to accomplish the matchbox problem with different numbers of matches in the two pockets.

```
function minmax=banachmultisim(N,p,sim)
if any(p<=0) | any(p>=1) | sum(p)~=1, error('bad p'), end
w=max(N); y=zeros(2,sim);
for i=1:sim
  [s,d]=banachmulti(N,p);
  y(1:2,i)=[s,d]';
end
tt1=tabulate(y(1,1:sim)+1); tt2=tabulate(y(2,1:sim)+1);
a=tt1(:,2); b=tt2(:,2); mx=tt1(end,1)-1; mi=tt2(end,1)-1;
a=a./sim; b=b./sim;
if length(N)==2
  plot(0:mx,a,'r-'), title('Mass Function of Remaining Matches')
else
  plot(0:mx,a,'r-',0:mi,b,'b--')
  title('Marginal Mass Function of Minimum and Maximum Remaining Matches')
end
minmax=y';

function [ma,mi]=banachmulti(x,p);
n=length(x); ok=1; mi=min(x);
while ok==1
  s=0; zaehler=0; r=unifrnd(0,1,1,1);
  for i=1:n, if zaehler==0
      s=s+p(i);
      if r<s, zaehler=i; end
  end, end
  x(zaehler)=x(zaehler)-1; mi=min(x); ma=max(x);
  if mi<1, ok=0; end
end
help=find(x==0); x(help)=max(x)+1; mi=min(x);
```

Program Listing 4: Code to accomplish the generalized matchbox problem.

Expectation $\mathbb{E}_N[K]$ is given by

$$\begin{aligned}
&\sum_{j=0}^{N} j \binom{2N-j}{N} \left(\frac{1}{2}\right)^{2N-j} = \sum_{k=0}^{N-1} (k+1) \binom{2N-k-1}{N} \left(\frac{1}{2}\right)^{2N-k-1} \\
&= \sum_{k=0}^{N-1} (k+1) \binom{2N-k-2}{N-1} \left(\frac{1}{2}\right)^{2N-k-1} + \sum_{k=0}^{N-1} (k+1) \binom{2N-k-2}{N} \left(\frac{1}{2}\right)^{2N-k-1} \\
&=: G + H.
\end{aligned}$$

Thus, G is given by

$$\begin{aligned}
&\frac{1}{2} \left\{ \sum_{k=0}^{N-1} k \binom{2(N-1)-k}{N-1} \left(\frac{1}{2}\right)^{2(N-1)-k} + \sum_{k=0}^{N-1} \binom{2(N-1)-k}{N-1} \left(\frac{1}{2}\right)^{2(N-1)-k} \right\} \\
&= \frac{1}{2} \mathbb{E}_{N-1}[K] + \frac{1}{2}
\end{aligned}$$

and, with $j = k+2$, H is

$$\begin{aligned}
&\sum_{j=2}^{N} (j-1) \binom{2N-j}{N} \left(\frac{1}{2}\right)^{2N-j+1} \\
&= \frac{1}{2} \left\{ \sum_{j=2}^{N} j \binom{2N-j}{N} \left(\frac{1}{2}\right)^{2N-j} - \sum_{j=2}^{N} \binom{2N-j}{N} \left(\frac{1}{2}\right)^{2N-j} \right\} \\
&= \frac{1}{2} \left\{ \mathbb{E}_N[K] - \binom{2N-1}{N} \left(\frac{1}{2}\right)^{2N-1} \right\} - \frac{1}{2} \left\{ 1 - \binom{2N}{N} \left(\frac{1}{2}\right)^{2N} - \binom{2N-1}{N} \left(\frac{1}{2}\right)^{2N-1} \right\} \\
&= \frac{1}{2} \mathbb{E}_N[K] - \frac{1}{2} + \binom{2N}{N} \left(\frac{1}{2}\right)^{2N+1}.
\end{aligned}$$

Simplifying $G + H$ yields the recursion

$$\mathbb{E}_N[K] = \mathbb{E}_{N-1}[K] + \binom{2N}{N} \left(\frac{1}{2}\right)^{2N}, \quad \mathbb{E}_1[K] = \frac{1}{2},$$

which resolves to

$$\mathbb{E}_N[K] = \sum_{i=1}^{N} \binom{2i}{i} \left(\frac{1}{2}\right)^{2i}.$$

This can be further simplified using several previous results, namely (1.29), (1.26), and (1.20), in that order. We list these three results, in that order, for convenience:

$$\begin{aligned}
\binom{2n}{n} &= (-1)^n 2^{2n} \binom{-\frac{1}{2}}{n}. \\
\binom{-n}{k} &= (-1)^k \binom{n+k-1}{k}. \\
\sum_{i=0}^{n} \binom{i+r-1}{i} &= \binom{n+r}{n}.
\end{aligned}$$

We then obtain, noting from (1.25) that $\binom{-1/2}{0} = 1$,

$$\begin{aligned}
\mathbb{E}_N[K] &= \sum_{i=1}^{N} \binom{2i}{i} \left(\frac{1}{2}\right)^{2i} \\
&= \sum_{i=1}^{N} (-1)^i \binom{-\frac{1}{2}}{i} \\
&= \sum_{i=1}^{N} \binom{i-1/2}{i} + 1 - 1 = \sum_{i=0}^{N} \binom{i-1/2}{i} - 1 \\
&= \binom{N+1/2}{N} - 1. \qquad (6.25)
\end{aligned}$$

This result can be extended to expressions for all moments of K. For example, some more work reveals that

$$\mathbb{V}(K) = [N - \mathbb{E}_N(K)] - \mathbb{E}_N(K)\,[1 + \mathbb{E}_N(K)] \approx \left(2 - \frac{4}{\pi}\right) N - \frac{2}{\sqrt{\pi}}\sqrt{N} + 2$$

and

$$\mathbb{E}_N\left(K^3\right) = 6\binom{N+\frac{3}{2}}{N} + 7\binom{N+\frac{1}{2}}{N} - 12N - 13,$$

with higher moments similarly obtained.[37]

Example 6.4 *We wish to show that the expectation (6.25) for Banach's matchbox problem in (6.24) can be well approximated by*

$$\mathbb{E}_N[K] \approx -1 + 2\sqrt{N/\pi}.$$

Let

$$w_n = \frac{2^n\, n!}{1 \cdot 3 \cdot 5 \cdots (2n-1)} = \frac{2 \cdot 4 \cdot 6 \cdots 2n}{1 \cdot 3 \cdot 5 \cdots (2n-1)} \approx \sqrt{n\pi}$$

from (2.236). Then, using the generalized binomial coefficient,

$$\begin{aligned}
\binom{N+1/2}{N} &= \frac{\left(N+\frac{1}{2}\right)\left(N-\frac{1}{2}\right)\cdots\frac{3}{2}}{N!} \\
&= \frac{(2N+1)\,(2N-1)\cdots 3}{2^N\, N!} = \frac{2N+1}{w_N} \approx \frac{2N+1}{\sqrt{n\pi}}. \qquad (6.26)
\end{aligned}$$

From Figure 41, we see that this approximation is very good even for modest values of N. Finally,

$$\mathbb{E}_N[K] \approx -1 + \frac{2N+1}{\sqrt{N\pi}} \approx -1 + 2\sqrt{N/\pi},$$

yielding the desired approximation. ■

[37] This example appears as Exercise 4.13 in Paolella, Fundamental Probability, and was kindly contributed by my friend and professor colleague Markus Haas.

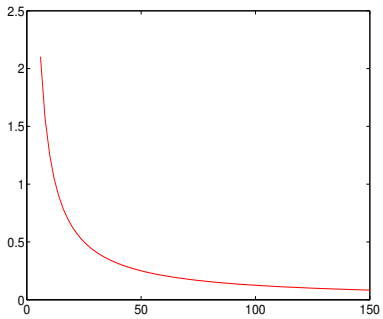


Figure 41: The relative percentage error, $100\,(\text{Approx} - \text{True})\,/\text{True}$, for the approximation in (6.26) as a function of N.