Lu Sun, and many more.

A Notebook on Calculus

To my family, friends and communities members who have been dedicating to the presentation of this notebook, and to all students, researchers and faculty members who might find this notebook helpful.

Contents

Foreword ix

Preface xi

List of Figures xiii

List of Tables xv

I Limit, Derivative and Integral 1

1 Limit **3**
- 1.1 Limit of Sequence 3
  - 1.1.1 Motivating Example 3
  - 1.1.2 Limit of Sequence 6
  - 1.1.3 Calculation of Limit of Sequence 8
- 1.2 Limit of Function 10
  - 1.2.1 Motivating Example 10
  - 1.2.2 Limit of a Function 11
  - 1.2.3 Calculation of the Limit of a Function 14

2 Derivative **17**
- 2.1 Motivating Example 17
- 2.2 Derivative of a Function 21
- 2.3 Calculation of the Derivative of a Function 21

3 Integral **23**
- 3.1 A Motivating Example 23
- 3.2 Integral of a Function 28
- 3.3 Calculation of the Integral of a Function 30

4 Applications **33**
- 4.1 Newton's Method 33
- 4.2 Taylor Series 34

II Multivariable Function 37

5 Multivariable Function **39**
5.1 Brief Introduction to Vector and Matrix 39
5.1.1 Basic Concepts 40
5.1.2 Matrix Multiplication 41
5.1.3 Block Matrix 42
5.1.4 Identity Matrix and Square Matrix Inverse 43
5.2 Multivariable Function 44

6 Partial derivative **45**
6.1 A Motivating Example 45
6.2 Partial Derivative 48
6.3 Gradient 51
6.3.1 Motivating Example 51
6.3.2 Gradient of a Multivariable Function 56
6.4 Jacobian Matrix 58

7 Multiple Integral **61**
7.1 Motivating Examples 61
7.2 Multiple Integral 66

8 Applications **69**
8.1 Neural Network Back-propagation 69
8.1.1 Perceptron 69
8.1.2 A Multi-layer Perceptrons Model 70
8.1.3 Training and Testing 70
8.2 Bayesian Inference 70

III Differential Equation **71**

9 Ordinary Differential Equation **73**
9.1 Definition 73
9.2 Canonical Forms and Solutions 73

10 Partial Differential Equation **75**
10.1 Definition 75
10.2 Canonical Forms and Solutions 75

11 Applications **77**
11.1 Circuit Transient Analysis 77
11.2 Finite Element Analysis 77

IV Calculus of Variations **79**

12 Functional **81**
- 12.1 Motivating Example 81
- 12.2 Functional 82
  - 12.2.1 Definition 82
  - 12.2.2 Finite Differences Approximation 83
  - 12.2.3 Function Space 84
  - 12.2.4 Normed Linear Space and Norm of Function 84
  - 12.2.5 Functional Continuity and Linearity 86
- 12.3 Variation of Functional 86

V Numerical Analysis **89**

13 Brief Introduction to Numerical Analysis **91**
- 13.1 Motivation 91
- 13.2 Challenges 92

14 One-Variable Equation **93**
- 14.1 General Problem Formulation 93
- 14.2 Solution 94
  - 14.2.1 Bisection Method 94
  - 14.2.2 Fixed Point Iteration Method 95
  - 14.2.3 Newton's method 97
  - 14.2.4 Secant Method 99
  - 14.2.5 Muller's Method 99
- 14.3 Convergence Speed and Error 99
  - 14.3.1 Order of Convergence 99
  - 14.3.2 Multiple Root 100
  - 14.3.3 Convergence Acceleration 101
- 14.4 Root for Polynomial 101

15 Interpolation **103**
- 15.1 Lagrange Interpolation 103

Bibliography **105**

Foreword

If software and e-books can be made completely open-source, why not a notebook?

This brings me back to the summer of 2009 when I started my third year as a high school student in Harbin No. 3 High School. In the end of August when the results of Gaokao (National College Entrance Examination of China, annually held in July) were released, people from photocopy shops would start selling notebooks photocopies that they claimed to be from the top scorers of the exam. Much curious as I was about what these notebooks look like, never have I expected myself to actually learn anything from them, mainly for the following three reasons.

First of all, some (in fact many) of these notebooks were more difficult to read than the textbooks. I guess we cannot blame the top scorers for being so smart that they sometimes made things extremely brief or overwhelmingly complicated.

Secondly, why would I want to adapt to notebooks of others when I had my own notebooks which in my opinion should be just as good as theirs.

And lastly, as a student in the top-tier high school myself, I knew that the top scorers were probably my schoolmates. Why would I pay money to a stranger in a photocopy shop for my friends' notebooks, rather than requesting a copy from them directly?

However, my mind changed after becoming an undergraduate student in 2010. There were so many modules and materials to learn for a college student, and as an unfortunate result, students were often distracted from digging deeply into a module (and for those who were still able to do so, you have my highest respect). The situation became worse when I started pursuing my Ph.D. in 2014. As I had to focus on specific research areas entirely, I could hardly split enough time on other irrelevant but still important and interesting contents.

To make a difference, I enforced myself reading articles beyond my comfort zone, which ended up motivating me to take notes to consolidate the knowledge. I used to work with hand-written notebooks. My very first notebook was on Numerical Analysis, an entrance-level module for engineering background graduate students. Till today I still have dozens of these notebooks on my bookshelf. Eventually, it came to me: why not digitizing them, making them accessible online and open-source and letting everyone read and edit it?

Similar with most open-source software, this notebook does not come with any "warranty" of any kind, meaning that there is no guarantee that everything in this notebook is correct, and it is not peer reviewed. **Do NOT cite this notebook in your academic research paper or book!** If you find anything helpful here with your research, please trace back to the origin of the knowledge and confirm by yourself.

This notebook is suitable as:

- a quick reference guide;
- a brief introduction for beginners of an area;
- a "cheat sheet" for students to prepare for the exam or for lecturers to prepare the teaching materials.

This notebook is NOT suitable as:

- a direct research reference;
- a replacement of the textbook.

The notebook is NOT peer reviewed, thus is more of a notebook than a book. It is meant to be easy to read, not to be comprehensive and very rigorous.

Although this notebook is open-source, the reference materials of this notebook, including textbooks, journal papers, conference proceedings, etc., may not be open-source. Very likely many of these reference materials are licensed or copyrighted. Please legitimately access these materials and properly use them, should you decided to trace the origin of the knowledge.

Some of the figures in this notebook are plotted using Excalidraw, a very convenient tool to emulate hand drawings. The Excalidraw project can be found on GitHub, *excalidraw/excalidraw*. Other figures may come from MATLAB, R, Python, and other computation engines. The source code to reproduce the results are intended to be included in the same repository of the notebook, but there might be exceptions.

This work might have benefited from the assistance of large language models, which are used exclusively for editing purposes such as correcting grammar and rephrasing sentences, without introducing new content, generating novel information, or changing the original intent of the text.

Preface

This notebook is on *Calculus*, a very important mathematical tool that was invented back in Newton's time or even earlier. It has now become an entrance level module for mathematics and engineering background students in their fist year in the university.

Initially, the invention of calculus, including the introduction of differentiation and integration, is of course used to explain things such as the concept of "speed" as a differential of distance over time. You might easily come up with common use cases of calculus, for example calculating the tangent of a curve, and calculating the volume of an arbitrarily shaped container. Other applications which may not make too much sense for beginners, for example the derivation of cycloid, are also done using calculus. Many advanced mathematical tools themselves are built on top of calculus, for example fourier transform, which is widely used in signal processing. Without a solid understanding of calculus, it is hardly possible for one to use these tools confidently and effectively.

The key references of this notebook is listed below. During the development of the notebook, this list may become longer and longer.

Book *Calculus Metric Version Eighth Edition* by James Stewart, published by Cengage Learning [2].

Book *Calculus* by Gibert Strang (Massachusetts Institute of Technology), published by Wellesley-Cambridge Press [3]. This book is available at MIT Open Courseware (*ocw.mit.edu*). There are countless number of great learning materials there.

List of Figures

1.1 Plot of a_n as a function of n in the motivating example. The readings of a_n are given in Table 1.2. 6
1.2 Plot of y as a function of x in the motivating example. 11
1.3 Plot of $y = \sin\left(\frac{1}{x}\right)$. 15

2.1 Plot of y as a function of x in the motivating example. 18
2.2 Slope of secant and tangent of $y = f(x)$ at $x = 2$. 19

3.1 Plot of (3.1) and $N = 3$ trapezoids. 24
3.2 Use $N = 20$ trapezoids to approximate the red area. 26
3.3 Calculation of the area of trapezoids. Variables $N = 3$ and $a = 0.5$ are used in the plot for demonstration. 27
3.4 Plot of $f(x) = \sin(x) + 0.5$ from 0 to 2π. The area above $y = 0$ is surrounded by the red dashed line, and the area below $y = 0$ by the blue dashed line. The definite integral $\int_0^{2\pi} f(x)dx$ in this case is the red area subtracting the blue area. 32

4.1 Plot of function $y = 2^x$ in red solid line, and its approximations using first-order, 5th-order and 10th-order Taylor series in blue solid line, blue dashed line and blue dot line respectively. . . 35

6.1 Plot of function $y = f(x_1, x_2)$ in 3-D. 46
6.2 The calculation of z using x and y. 49
6.3 Plot of $y = f(x_1, x_2)$ in 3-D. 52
6.4 Contour line of $y = f(x_1, x_2)$. 53
6.5 Plot of vectors given by (6.16) and (6.17). 54
6.6 Formulation of the tangent plane from vectors given by (6.16) and (6.17). 54
6.7 Trajectory of x until the maximum $y = f(x)$ is achieved. . . . 57
6.8 Trajectory of x until the maximum $y = f(x)$ is achieved on contour line plot. 57

7.1 Calculation of the volume of a cone. 62
7.2 Calculation of volume of an arbitrary arbitrary project. . . . 64
7.3 Calculation of volume of an arbitrary object as sum of cubes. 65
7.4 Summation of cubes in different sequence. 66

12.1 Brachistochrone problem description. 82

14.1 A demonstrative example of solving $f(x) = 0$ using numerical methods. 94
14.2 Function $g(x) = x^2 - 2$ and its two fixed points, $p_1 = -1$ and $p_2 = 2$. 96
14.3 Approaching fixed point using fixed point iteration. 97
14.4 A demonstration of using Newton's method to solve an equation. 98

List of Tables

1.1 Calculate a_n for any arbitrary n in the motivating example. . 4
1.2 Calculate a_n for any arbitrary n in the motivating example, but longer table. 5
1.3 Convergence of commonly seen $\{a_n\}$ and $\{s_n\}$. Variable c, r are constants. 8
1.4 Limit of commonly seen elementary functions. 14

2.1 Derivative of commonly seen functions. 22

3.1 Indefinite integral of commonly seen functions. 30

Part I

Limit, Derivative and Integral

1

Limit

CONTENTS

1.1 Limit of Sequence 3
  1.1.1 Motivating Example 3
  1.1.2 Limit of Sequence 6
  1.1.3 Calculation of Limit of Sequence 8
1.2 Limit of Function 10
  1.2.1 Motivating Example 10
  1.2.2 Limit of a Function 11
  1.2.3 Calculation of the Limit of a Function 13

Consider a sequence $\{a_n\}$ or a function $f(x)$. In this chapter, the values of a_n or $f(x)$ when n or x grows towards infinity are discussed. The value of $f(x)$ when x approaches a constant value is discussed.

1.1 Limit of Sequence

A motivating example of the limit of a sequence is given in Section 1.1.1. The definition of the limit of a sequence is given in Section 1.1.2. The calculation of the limit of a sequence is discussed in 1.1.3, mainly on the proof of convergence of a sequence.

1.1.1 Motivating Example

We use $\{a_n\}$ to denote a **sequence**. In $\{a_n\}$, the positive integer n is the index of the elements in the sequence, where a_1 represents the first element of $\{a_n\}$, and a_2 the second element, and so on. A sequence has at least one element a_1. It may have finite elements, in which case it is called a **finite sequence**. Or, it may have infinite elements, in which case it is called an **infinite sequence**. In this notebook, we are mostly interested in infinite sequences.

A motivating example is given below to illustrate the limit of an infinite sequence.

A Motivating Example

Consider an infinite sequence $\{a_n\}$ whose elements are recursively calculated by

$$a_1 = 1, \tag{1.1}$$

$$a_n = a_{n-1} + \left(\frac{1}{2}\right)^{n-1}. \tag{1.2}$$

Q1: Calculate the feasible domain of n such that $a_n \geq 1.95$.
Q2: Calculate the feasible domain of n such that $a_n \geq 2$.

The old school way of solving the questions is rather simple: use a table to list down different n and its corresponding a_n. The value of a_n can be manually calculated for small n, as shown in Table 1.1. In theory, this table can be extended to any n of choice. From Table 1.1, for any $n \geq 6$, $a_n \geq 1.95$.

n	$a_n - a_{n-1} = \left(\frac{1}{2}\right)^{n-1}$	a_n
1	—	1
2	0.5	1.5
3	0.25	1.75
4	0.125	1.875
5	0.0625	1.9375
6	0.03125	1.96875
7	0.015625	1.984375
⋮	⋮	⋮

To find out the feasible range of n such that $a_n \geq 2$, we can start by investigating a larger table, say Table 1.2. From Table 1.2, it can been seen that as n grows larger and larger, the increment $a_n - a_{n-1} = \left(\frac{1}{2}\right)^{n-1}$ becomes smaller and smaller, and the increment is just hardly enough to top a_n to 2.

An alternative method to find the feasible range is to derive an analytical equation of a_n as a function of n, and we can solve $a_n \geq 2$ for n. Recursively using (1.2) for $t - 1$ times and substituting (1.1) into (1.2) give

$$a_n = \sum_{i=1}^{n} \left(\frac{1}{2}\right)^{i-1} \tag{1.3}$$

$$= 1 + \sum_{i=2}^{n} \left(\frac{1}{2}\right)^{i-1}, \tag{1.4}$$

n	$a_n - a_{n-1} = \left(\frac{1}{2}\right)^{n-1}$	a_n
1	1	1
2	0.5	1.5
3	0.25	1.75
4	0.125	1.875
5	0.0625	1.9375
6	0.03125	1.96875
7	0.015625	1.984375
8	0.0078125	1.9921875
9	0.00390625	1.99609375
10	0.001953125	1.998046875
11	0.0009765625	1.9990234375
12	0.00048828125	1.99951171875
13	0.000244140625	1.999755859375
14	0.0001220703125	1.9998779296875
15	0.00006103515625	1.99993896484375
$\vdots$	$\vdots$	$\vdots$

and from (1.3)

$$\begin{aligned} \frac{1}{2}a_n &= \sum_{i=1}^{n}\left(\frac{1}{2}\right)^{i} \\ &= \sum_{i=2}^{n}\left(\frac{1}{2}\right)^{i-1} + \left(\frac{1}{2}\right)^{n}. \end{aligned} \tag{1.5}$$

Subtracting (1.5) from (1.4) gives

$$\begin{aligned} \frac{1}{2}a_n &= 1 - \left(\frac{1}{2}\right)^{n}, \\ a_n &= 2 - \left(\frac{1}{2}\right)^{n-1}. \end{aligned} \tag{1.6}$$

Equation (1.6) can be verified using the results in Table 1.2. It suggests that the elements in sequence $\{a_n\}$ will never reach 2 for any n, although a_n can be very close to 2 as n increases. This can also be shown from the plot of (1.6) as given in Fig. 1.1.

The following features can be observed for sequence $\{a_n\}$ from both (1.6) and Fig. 1.1.

- The sequence is monotonically increasing, as $a_{n-1} < a_n$ for any n.
- The sequence is bounded, as $a_n < 2$ for any n.

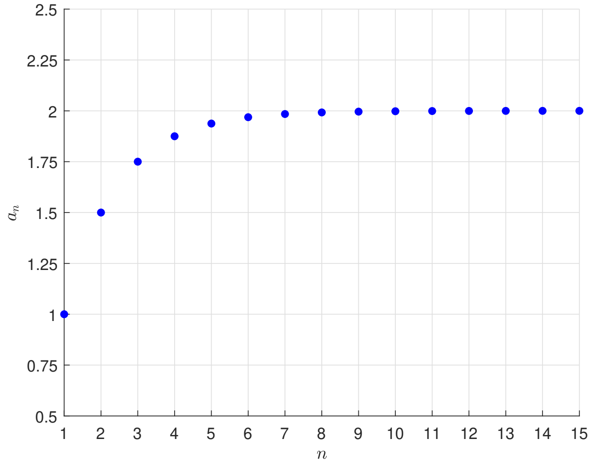

FIGURE 1.1
Plot of a_n as a function of n in the motivating example. The readings of a_n are given in Table 1.2.

- The sequence can get "as close to 2 as we like", in the sense that for any value smaller than 2, however close to 2 it is (say, 1.9999), a_n will at some point exceeds that value and get even closer to 2, for large enough n.

Sequence $\{a_n\}$ reveals an important but not intuitive fact: it is possible to find a monotonically increasing yet bounded sequence. Another way to look at it is that it is possible to add infinite number of positive values together, yet the result is a finite value. From the fact that $\{a_n\}$ gets as close to a certain value as possible when n gets large enough, the limit of sequence is defined, as introduced in the next Section 1.1.2.

1.1.2 Limit of Sequence

Given an infinite sequence $\{a_n\}$, if $\{a_n\}$ is bounded and it gets close to a value L "as close as we like" for large n, then L is called the **limit of the sequence**. The formal definition of the limit of a sequence is given below.

Definition of the limit of sequence:

A sequence $\{a_n\}$ has limit L if for any $\varepsilon > 0$ (however small it is), there is always a corresponding integer N, such that if

$n > N$, $|a_n - L| < \varepsilon$. This is denoted by

$$\lim_{n\to\infty} a_n \quad = \quad L,$$

or

$$a_n \to L \quad \text{as} \quad n \to \infty,$$

and in this case we say "sequence $\{a_n\}$ is convergent" and "sequence $\{a_n\}$ converges to L (as n approaches infinity)".

If an infinite sequence has a limit, it is known as a **convergent sequence**. Otherwise, it is known as a **divergent sequence**. The sequence $\{a_n\}$ in (1.6) is an example of a convergent sequence that converges to 2. We can easily prove this using the definition of the limit of a sequence as follows.

Given any $\varepsilon > 0$, from (1.6) solving

$$|a_n - 2| \quad < \quad \varepsilon$$

gives

$$\begin{aligned} \left| 2 - \left(\frac{1}{2}\right)^{n-1} - 2 \right| \quad &< \quad \varepsilon, \\ n \quad &> \quad 1 - \log_2 \varepsilon. \end{aligned} \tag{1.7}$$

As long as N is an integer satisfying $N > 1 - \log_2 \varepsilon$, for $n > N$, $|a_n - 2| < \varepsilon$ can be achieved. For example, if $\varepsilon = 0.05$, (1.7) gives $n > 5.32$. This implies that by letting $N = 6$ (or 7, 8, ...), for any $n \geq N$, $|a_n - 2| < 0.05$. This matches the observation given in Table 1.1.

If there is no such limit L for an infinite sequence $\{a_n\}$, we say "sequence $\{a_n\}$ is divergent" or "sequence $\{a_n\}$ diverges".

As a special case of divergent sequences, if $\{a_n\}$ becomes unbounded as n approaches infinity, we say "sequence $\{a_n\}$ diverges to infinity". The definition is as follows.

Definition of sequence diverging to infinity:

For a sequence $\{a_n\}$, if for any arbitrary positive value M (however large it is), there is always a corresponding integer N, such that if $n > N$, $|a_n| > M$, we say "$\{a_n\}$ diverges to infinity". This is denote it by

$$\lim_{n\to\infty} a_n \quad = \quad \infty.$$

Sequence $\{s_n\}$ where $s_n = \sum_{i=1}^{n} a_n$ is used to denote the sum of the first n elements in the infinite sequence $\{a_n\}$. Apparently, for infinite sequence $\{a_n\}$, $\{s_n\}$ is also an infinite sequence which may or may not converge depending on $\{a_n\}$. The limit of $\{s_n\}$ is denoted by

$$\lim_{n\to\infty} s_n = \sum_{i=1}^{\infty} a_n$$

if $\{s_n\}$ converges and $\sum_{i=1}^{\infty} a_n$ exists. Otherwise, $\{s_n\}$ diverges and $\sum_{i=1}^{\infty} a_n$ does not exist.

1.1.3 Calculation of Limit of Sequence

As the first step in the calculation of the limit of a sequence, the convergence of the sequence must be proved. It makes no sense to calculate the limit if it does not exist in the first place. This section mainly focuses on the proof of convergence of a sequence. Usually, after proving the convergence, the limit can be obtained by either using numerical methods or calculating the limit of the function $a_n = f(n)$ as a function of n. More details of calculating the limit of a function are given in Section 1.2.

Generally speaking, there is no systematic way of proving the convergence or divergence of an infinite sequence except using the definition. Sometimes the proof can be very difficult or even impossible. Some well-studied and commonly seen sequences are summarized in the following Table 1.3.

TABLE 1.3
Convergence of commonly seen $\{a_n\}$ and $\{s_n\}$. Variable c, r are constants.

Category	a_n	s_n	$\lim_{n\to\infty} a_n$	$\lim_{n\to\infty} s_n$
Polynomial	c, $c \neq 0$	nc	c	∞
	n	$\frac{n(n+1)}{2}$	∞	∞
	n^2	$\frac{n(n+1)(2n+1)}{6}$	∞	∞
	n^3	$\frac{n^2(n+1)^2}{4}$	∞	∞
Power Series	cr^{n-1}, $\lvert r\rvert< 1$	$\frac{c(1-r^n)}{1-r}$	0	$\frac{c}{1-r}$
Others	n^{-1}	—	0	∞
	n^{-2}	—	0	$\frac{\pi}{6}$

"∞" stands for "diverges to infinity".

If a sequence falls into one of the above categories, or somewhat similar to the above categories, or can be expressed as a compound sequence derived from the above categories, its convergence or divergence might be proved a bit easier. For example, if $c_n = a_n + b_n$, and both a_n and b_n converge, then

$$\lim_{n\to\infty} c_n = \lim_{n\to\infty} a_n + \lim_{n\to\infty} b_n.$$

This can be proved using the definition of the limitation of the sequence.

Some other interesting features regarding sequence convergence are given as follows. Given two sequences $\{a_n\}$ and $\{b_n\}$, if

$$\lim_{n\to\infty} \frac{a_n}{b_n} = L \neq 0,$$

then $\{a_n\}$ and $\{b_n\}$ must behave the same in terms of convergence, meaning that both of them must converge or diverge at the same time.

It is intuitive and not difficult to prove that for $\{s_n\}$ to converge, i.e. $\lim_{n\to\infty} s_n = s$, it is necessary (but not sufficient) for its corresponding $\{a_n\}$ to converge to zero, i.e. $\lim_{n\to\infty} a_n = 0$. Do notice that it is possible to have a divergent $\{s_n\}$ even if a_n converges to zero. An example is the harmonic series given in Table 1.3 where $a_n = n^{-1}$.

The famous **monotone convergence theorem for sequence** given below can become handy when proving the convergence of a sequence.

Monotone Convergence Theorem for Sequence:

If a sequence $\{a_n\}$ is monotonically increasing or decreasing, and meantime it is bounded, i.e. $|a_n| < M$ for all n, then $\{a_n\}$ must be convergent

---

The proof of this theorem is not included in this notebook.

From the monotone convergence theorem, we know that for a sequence $\{a_n\}$, if $\sum_{i=1}^{\infty}|a_n|$ exists, then the infinite sum $\sum_{i=1}^{\infty} a_n$ must also exist. This can be illustrated simply by splitting $\{a_n\}$ into $\{a_n^+\}$ and $\{a_n^-\}$, where

$$\begin{aligned} a_n^+ &= \begin{cases} a_n & a_n \geq 0 \\ 0 & a_n < 0 \end{cases}, \\ a_n^- &= \begin{cases} -a_n & a_n < 0 \\ 0 & a_n \geq 0 \end{cases}. \end{aligned}$$

Apparently, both a_n^+ and a_n^- are non-negative. Therefore, both $\sum_n a_n^+$ and $\sum_n a_n^-$ are monotonically increasing non-negative sequences. We also know that both of them are bounded because $\sum_n |a_n| = \sum_n a_n^+ + \sum_n a_n^-$ is bounded. Therefore, both $\sum_n a_n^+$ and $\sum_n a_n^-$ must be convergent according to the monotone convergence theorem. This implies that $\sum a_n = \sum a_n^+ - \sum a_n^-$ must also be a convergent sequence as it is the sum of two convergent sequences.

Sequence a_n with bounded $\sum_{i=1}^{\infty}|a_n|$ is known as an **absolutely convergent sequence**. An absolutely convergent sequence a_n must have a bounded infinite sum $\sum_{i=1}^{\infty} a_n$, but might not be true wise versa.

1.2 Limit of Function

A motivating example is given in Section 1.2.1. The definition of the limit of a function is given in Section 1.2.2. The calculation of the limit of a function is discussed in 1.2.3.

1.2.1 Motivating Example

A motivating example is given below to illustrate the limit of a function.

A Motivating Example

Consider function

$$y = f(x) = \begin{cases} (x-1)^2 & x \neq 1 \\ 1 & x = 1 \end{cases} . \tag{1.8}$$

Q1: Obtain the domain for x and range for y.
Q2: Calculate y at $x = 1$.
Q3: Calculate y when x is in a small "neighbourhood" of 1, but $x \neq 1$.

For Q1, the plot of y as a function of x is given in Fig 1.2. It is clear from the figure that the domain and range of the function are $x \in \mathbb{R}$ and $y \in \mathbb{R}, y > 0$ respectively. For Q2, substituting $x = 1$ into (1.8) gives $y = 1$, as it is also shown in Fig 1.2.

For Q3, we need to rewrite "x is in a small neighbourhood of 1, but $x \neq 1$" in a bit more precise manner. An intuitive way is to define a small "threshold area" near $x = 1$, say, $-\delta < x - 1 < \delta, x \neq 1$ with δ being a very small positive value. Notice that $x = 1$ is not concerned and the value of y near $x = 1$ has nothing to do with y at $x = 1$.

Next, we can calculate y subject to $-\delta < x - 1 < \delta, x \neq 1$. Clearly, the range of y relates to the choice of δ. Substituting $-\delta < x - 1 < \delta, x \neq 1$ into (1.8) gives $0 < y < \delta^2$. With δ being chosen smaller and smaller, the range of y would become smaller and smaller, and y will eventually approach 0 (although y cannot be precisely 0).

This discussion is essentially a study of y when x is "as close as we like but not equal" to 1. We have learned that in this motivating example:

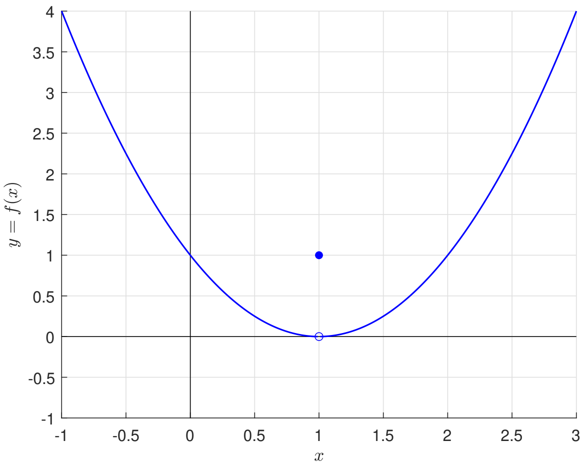

FIGURE 1.2
Plot of y as a function of x in the motivating example.

- The value of y when x is "as close as we like to 1" does not rely on the value of y at $x = 1$.
- The value of y when x is "as close as we like to 1" floats in a small range depending on the size of the "neighbourhood" of 1 where x residents. The size of the "neighbourhood" is quantified by δ in this example. Thus, the range of y is related to the choice of δ.
- With a proper choice of δ, the value of y can be as "close as we like to 0".

The above features give a brief idea of limit of a function. In the example, y can be made as close as we like to 0 simply by making x close enough to 1 by choosing a small δ.

1.2.2 Limit of a Function

The formal definition of the **limit of the function** is given below. Notice that there are a few different but equivalent ways to define the limit of a function. Here the "ε-δ definition" is introduced.

Definition of the limit of the function at $x \to a$:

A function $f(x)$ of x has the limit L at $x = a$ if for any $\varepsilon > 0$, there is always a corresponding $\delta > 0$, such that if $0 < |x-a| < \delta$,

$|f(x)-L|<\varepsilon$, with the prerequisite that $0<|x-a|<\delta$ is defined for $f(x)$. This is denoted by

$$\lim_{x\to a} f(x) \quad = \quad L,$$

or

$$f(x)\to L \quad \text{as} \quad x\to a.$$

Using the definition above, it can be proved easily that for the motivating example in Section 1.2.1, the function $y=f(x)$ has a limit of $\lim_{x\to 1} f(x)=0$. Notice that $\lim_{x\to a} f(x)=L$ does not necessarily require $f(a)=L$. As a matter of fact, $f(x)$ does not even need to be defined at $x=a$, as long as it is defined at the neighbour of $x=a$.

Similar to the definition of the limit of a function, the definition of **one-sided limit of the function** is given below. The one-sided limit is similar but weaker than the definition of the limit of a function in the sense that it only concerns one side of the neighbour of $x=a$.

Definition of the one-sided limit of the function:

A function $f(x)$ of x has the one-side left limit L^- at $x=a$ if for any $\varepsilon>0$, there is always a corresponding $\delta>0$, such that if $a-\delta<x<a$, $|f(x)-L^-|<\varepsilon$, with prerequisite that $a-\delta<x<a$ is defined for $f(x)$. This is denoted by

$$\lim_{x\to a^-} f(x) \quad = \quad L^-,$$

or

$$f(x)\to L^- \quad \text{as} \quad x\to a^-.$$

A function $f(x)$ of x has the one-side right limit L^+ at $x=a$ if for any $\varepsilon>0$, there is always a corresponding $\delta>0$, such that if $a<x<a+\delta$, $|f(x)-L^+|<\varepsilon$, with prerequisite that $a<x<a+\delta$ is defined for $f(x)$. This is denoted by

$$\lim_{x\to a^+} f(x) \quad = \quad L^+,$$

or

$$f(x)\to L^+ \quad \text{as} \quad x\to a^+.$$

It is clear from the definition that a function $f(x)$ has a limit of L at $x = a$ if and only if it has both one-sided left limit L^- and one-sided right limit L^+ at $x = a$ and $L^- = L^+ = L$, i.e.

$$\lim_{x\to a} f(x) = L \quad \Leftrightarrow \quad \lim_{x\to a^-} f(x) = \lim_{x\to a^+} f(x) = L.$$

Furthermore, if function $f(x)$ has a limit L at $x = a$, and also $f(x) = L$, the function $f(x)$ is called a **continuous function at** $x = a$. The example given in the motivating example in Section 1.2.1 is not continuous at $x = 1$ as $\lim_{x\to 1} = 0$ while $f(x)|_{x=1} = 1$, which can be seen from Fig. 1.2. However, it is continuous everywhere else. For instance, at $x = 0$, $\lim_{x\to 0} = 1$ and $f(x)|_{x=0} = 1$.

The definition of the limit of a function $f(x)$ when x approaches infinity is given below. It is quite similar to the definition of the limit of an infinite sequence.

Definition of the limit of the function at $x \to \pm\infty$:

A function $f(x)$ of x has the limit L at $x \to +\infty$ (sometimes denoted as $x \to \infty$ for simplicity) if for any $\varepsilon > 0$, there is always a corresponding δ, such that if $x > \delta$, $|f(x) - L| < \varepsilon$, with prerequisite that $x > \delta$ is defined for $f(x)$. This is denoted by

$$\lim_{x\to +\infty} f(x) \quad = \quad L,$$

or

$$f(x) \to L \quad \text{as} \quad x \to +\infty.$$

A function $f(x)$ of x has the limit L at $x \to -\infty$ if for any $\varepsilon > 0$, there is always a corresponding δ, such that if $x < \delta$, $|f(x) - L| < \varepsilon$, with prerequisite that $x < \delta$ is defined for $f(x)$. This is denoted by

$$\lim_{x\to -\infty} f(x) \quad = \quad L,$$

or

$$f(x) \to L \quad \text{as} \quad x \to -\infty.$$

In special cases where the function is unbounded, the limit of the function does not exist and we can use $\lim_{x\to a} f(x) = \infty$ and $\lim_{x\to\pm\infty} f(x) = \infty$ to represent the cases.

1.2.3 Calculation of the Limit of a Function

The calculation of the limit of many commonly seen elementary functions are often obvious and easy. This is because these functions are often continuous in the domain, and for a continuous function the limit $\lim_{x\to a} f(x)$ can be obtained by simply substituting $x = a$ into the function. The limit $\lim_{x\to\infty} f(x)$ might be slightly difficult but mostly can be obtained from the definition.

Some examples are given below in Table 1.4. For the case of rational function, if $p(a) \neq 0$, $\lim_{x\to a} \dfrac{q(x)}{p(x)} = \dfrac{q(a)}{p(a)}$. If $p(a) = 0$, $\lim_{x\to a} \dfrac{q(x)}{p(x)}$ depends on the order and coefficients of $q(x)$ and $p(x)$ and may not exist. The limit $\lim_{x\to\infty} \dfrac{q(x)}{p(x)}$ depends on the order and coefficients of $q(x)$ and $p(x)$ and may not exist.

TABLE 1.4
Limit of commonly seen elementary functions.

Category	$f(x)$	$\lim_{x\to a} f(x)$	$\lim_{x\to\infty} f(x)$
Polynomial	$p(x)$	$p(a)$	∞
Root	$\sqrt{x}$	$\sqrt{a}$ for $a > 0$	∞
Rational	$\dfrac{q(x)}{p(x)}$	depends	depends
Trigonometric	$\sin(x)$, $\cos(x)$	$\sin(a)$, $\cos(a)$	No
Exponential	e^{-x}	e^{-a}	0 as $x \to +\infty$ ∞ as $x \to -\infty$
Logarithm	$\log_e(x)$	$\log_e(a)$ for $a > 0$	∞

"∞" stands for "unbounded".

It is worth mentioning a few typical cases where a function $f(x)$ does not have a limit and/or is not continuous.

Case 1: function $f(x)$ does not converge at a neighbourhood of $x = a$, thus does not have a one-sided limit. For example, consider

$$f(x) = \sin\left(\frac{1}{x}\right).$$

The function is defined at $x \in \mathbb{R}, x \neq 0$, but it does not have a one-sided limit $\lim_{x\to 0^-} f(x)$ or $\lim_{x\to 0^+} f(x)$, as when $x \to 0$, function $f(x)$ oscillates, as shown in Fig. 1.3.

Case 2: function $f(x)$ is unbounded at a neighbourhood of $x = a$, therefore does not have a one-sided limit. For example, consider

$$f(x) = \left|\frac{1}{x}\right|.$$

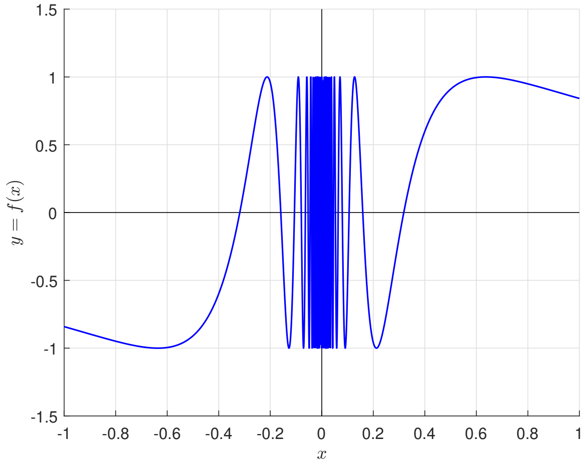

FIGURE 1.3
Plot of $y = \sin\left(\frac{1}{x}\right)$.

Apparently, $\lim_{x\to 0^-} f(x)$ or $\lim_{x\to 0^+} f(x)$ does not exist.

Case 3: function $f(x)$ has one-sided limits $\lim_{x\to 0^-} f(x)$ and $\lim_{x\to 0^+} f(x)$, but $\lim_{x\to 0^-} f(x) \neq \lim_{x\to 0^+} f(x)$. For example, consider

$$f(x) \quad = \quad \mathrm{sign}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}.$$

From the definition, $\lim_{x\to 0^-} f(x) = -1$ and $\lim_{x\to 0^+} f(x) = 1$, therefore, $\lim_{x\to 0} f(x)$ does not exist.

Some commonly used tricks are as follows. For $g(x) = f_1(x) + f_2(x)$, if the limits exist for $f_1(x)$ and $f_2(x)$ at $x \to a$, then

$$\lim_{x\to a} g(x) \quad = \quad \lim_{x\to a} f_1(x) + \lim_{x\to a} f_2(x).$$

The same is true for the cases $g(x) = f_1(x)f_2(x)$ and $g(x) = \frac{f_1(x)}{f_2(x)}$, subject to $\lim_{x\to a} f_2(x) \neq 0$ when it is the denominator. And it holds true for $x \to \pm\infty$ as well.

In the case of $g(x) = \frac{f_1(x)}{f_2(x)}$ and $\lim_{x\to a} f_2(x) = 0$, the discussion is more complicated. Sometimes, *L'Hôpital's rule* can become handy.

L'Hôpital's rule:

For $g(x) = \dfrac{f_1(x)}{f_2(x)}$

$$g(x) \quad = \quad \frac{f_1(x)}{f_2(x)}$$

if the following criteria are satisfied:

- Both $\lim_{x\to a} f_1(x) = 0$ and $\lim_{x\to a} f_2(x) = 0$.
- Both $f_1(x)$, $f_2(x)$ are continuous and differentiable at $x = a$.
- The derivative of the denominator $\lim_{x\to a} f_2'(x) \neq 0$.

then $\lim_{x\to a} g(x)$ can be calculated using (1.9), provided that the right side of (1.9) exists.

$$\lim_{x\to a} g(x) = \lim_{x\to a} \frac{f_1'(x)}{f_2'(x)}. \tag{1.9}$$

where $f'(x)$ is the derivative of $f(x)$.

The derivative of a function $f(x)$ will be introduced in the next Chapter 2. The proof of L′Hôpital's rule requires solid calculus foundation and is out of the scope of this notebook.

The limit of an infinite sequence is linked to the limit of the associated function at $x \to \infty$. For example, for sequence $\{a_n\}$, if $a_n = f(n)$ and $\lim_{n\to\infty} f(n) = L$, then $\lim_{n\to\infty} a_n = L$.

2

Derivative

CONTENTS

2.1 Motivating Example 17
2.2 Derivative of a Function 20
2.3 Calculation of the Derivative of a Function 21

Consider a continuous function $f(x)$ defined in $[a, b]$. In this chapter the ratio of change of $f(x)$ versus x is quantitively derived and studied.

2.1 Motivating Example

Consider the following motivating example.

A Motivating Example

Consider

$$y = f(x) = (|x|-1)^2 . \tag{2.1}$$

Obviously the above function (2.1) is continuous in $x \in \mathbb{R}$. We want to study the change $\Delta y = f(x + \Delta x) - f(x)$ given a small deviation Δx at different values of x.

Q1: Describe Δy given a small deviation Δx at $x = 2$, and describe the ratio $\frac{\Delta y}{\Delta x}$ when $\Delta x \to 0$.

Q2: Describe Δy given a small deviation Δx at $x = 0$, and describe the ratio $\frac{\Delta y}{\Delta x}$ when $\Delta x \to 0$.

Q3: Describe the ratio $\frac{\Delta y}{\Delta x}$ at any x when $\Delta x \to 0$.

As a first step, plot y as a function of x from (2.1) in Fig. 2.1 for convenient analysis.

Consider $x = 2$. Variable $\Delta y = f(x + \Delta x) - f(x)$ is given in (2.2). Notice that the deviation Δx is supposed to be small, i.e. $|\Delta x| \approx 0$. Therefore, we can

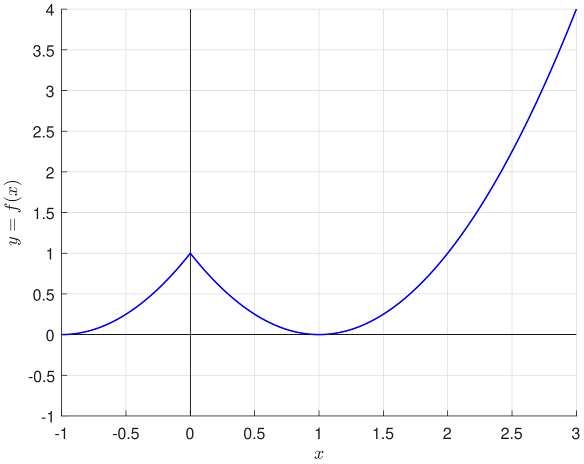

FIGURE 2.1
Plot of y as a function of x in the motivating example.

safely assume $2 + \Delta x \geq 0$ for simplicity.

$$\Delta y = f\left(2 + \Delta x\right) - f(2) = \left(2 + \Delta x - 1\right)^2 - \left(2 - 1\right)^2 = 2\Delta x + \Delta x^2. \quad (2.2)$$

With (2.2), it is possible to calculate y at $x = 2+\Delta x$ by using $y = f(2)+\Delta y$ for small Δx. For example, to calculate $f(1.9)$, substituting $\Delta x = -0.1$ into (2.2) gives $\Delta y = -0.19$, thus $f(1.9) = f(2) - 0.19 = 0.81$.

From (2.2), the ratio $\frac{\Delta y}{\Delta x}$ can be calculated as

$$\left.\frac{\Delta y}{\Delta x}\right|_{x=2} = 2 + \Delta x, \quad (2.3)$$

where $(\cdot)|_{x=a}$ represents substituting $x = a$ into $(\cdot)$.

Notice that (2.3) can be interpreted geometrically as the slope of secant of $(2, 1)$ and $(x + \Delta x, f\left(x + \Delta x\right))$. For example, when $\Delta x = 0.5$, the slope of secant $(2, 1)$ and $(2.5, 2.25)$ can be obtained using (2.3) as 2.5, which is shown as the red solid line in Fig. 2.2. Other two examples when $\Delta x = 0.9$ and $\Delta x = -0.5$ are given by red dashed line and red dot-dashed line in Fig. 2.2 respectively, and their slopes can be calculated as 2.9 and 1.5 respectively using (2.3).

Apparently, the slope of the secant depends on Δx, which can be seen from both equation (2.3) and Fig. 2.2. From the figure, when $\Delta x \to 0$, the slope of the tangent at $x = 2$ can be obtained, as given by the blue dashed line in Fig. 2.2. From (2.3),

$$\lim_{\Delta x \to 0} \left.\frac{\Delta y}{\Delta x}\right|_{x=2} = 2 \quad (2.4)$$

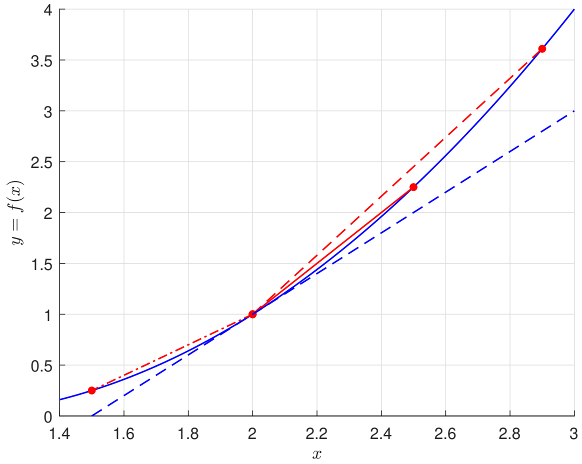

FIGURE 2.2
Slope of secant and tangent of $y = f(x)$ at $x = 2$.

which is the slope of the tangent of $y = f(x)$ at $x = 2$.

Consider $x = 0$. Variable $\Delta y = f(x + \Delta x) - f(x)$ can be obtained as given in (2.5). Different from (2.2), this time the analytical equation has two forms, depending on the sign of Δx.

$$\Delta y = f(\Delta x) - f(0) = \begin{cases} \Delta x^2 - 2\Delta x & \Delta x > 0 \\ \Delta x^2 + 2\Delta x & \Delta x < 0 \end{cases}. \tag{2.5}$$

From (2.5), the ratio $\frac{\Delta y}{\Delta x}$ can be calculated as

$$\left.\frac{\Delta y}{\Delta x}\right|_{x=0} = \begin{cases} -2 + \Delta x & \Delta x > 0 \\ 2 + \Delta x & \Delta x < 0 \end{cases}. \tag{2.6}$$

Equation (2.6) implies that

$$\begin{aligned} \lim_{\Delta x \to 0^-} \left.\frac{\Delta y}{\Delta x}\right|_{x=0} &= 2, \\ \lim_{\Delta x \to 0^+} \left.\frac{\Delta y}{\Delta x}\right|_{x=0} &= -2, \end{aligned}$$

and $\lim_{\Delta x \to 0} \left.\frac{\Delta y}{\Delta x}\right|_{x=0}$ does not exist. This can be intuitively comprehended from Fig. 2.1 as there is no tangent for the curve at $x = 0$.

Consider Q3. From what have been achieved in Q1 and Q2, we know that at different value of x, the ratio $\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$ can be different. In this example, we need to discuss three sub-cases $x < 0$, $x = 0$ and $x > 0$. Notice that since

Δx is small ($\Delta x \to 0$ is studied), we assume $x + \Delta x$ has the same sign with x when $x \neq 0$.

When $x < 0$,

$$\begin{aligned} \Delta y &= f\left(x + \Delta x\right) - f(x) \\ &= \left(-x - \Delta x - 1\right)^2 - \left(-x - 1\right)^2 \\ &= \Delta x^2 + 2x\Delta x + 2\Delta x, \end{aligned}$$

and the ratio is

$$\frac{\Delta y}{\Delta x} = \Delta x + 2x + 2.$$

Thus,

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = 2x + 2 \tag{2.7}$$

When $x = 0$, as discussed previously, $\lim_{x \to 0} \frac{\Delta y}{\Delta x}$ does not exist.

When $x > 0$,

$$\begin{aligned} \Delta y &= f\left(x + \Delta x\right) - f(x) \\ &= \left(x + \Delta x - 1\right)^2 - \left(x - 1\right)^2 \\ &= \Delta x^2 + 2x\Delta x - 2\Delta x, \end{aligned}$$

and the ratio is

$$\frac{\Delta y}{\Delta x} = \Delta x + 2x - 2.$$

Thus,

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = 2x - 2 \tag{2.8}$$

To summarize (2.7) and (2.8)

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \left\{ \begin{array}{ll} 2x + 2 & x < 0 \\ 2x - 2 & x > 0 \end{array} \right. , \tag{2.9}$$

which can be interpreted as the slope of tangent of curve $y = f(x)$ at different x in Fig. 2.1.

This motivating example shows that for a function $y = f(x)$, the ratio $\lim_{\Delta x \to 0} \frac{f(x+\Delta x)}{\Delta x}$ may change for different x, and sometimes may not even exist. The limit $\lim_{\Delta x \to 0} \frac{f(x+\Delta x)}{\Delta x}$ itself is also a function of x, which we call the **derivative** of $f(x)$.

2.2 Derivative of a Function

The formal definition of the derivative of a scalar function $f(x)$ with respect to x is given as follows.

Definition of the derivative of a function:

The derivative of $f(x)$ at $x = a$, denoted by $f'(a)$, is given as follows

$$f'(a) = \lim_{\Delta x \to 0} \frac{f(a + \Delta x) - f(a)}{\Delta x} \tag{2.10}$$

if such limit in (2.10) exists. In this case, $f(x)$ is called differentiable at $x = a$.

If $y = f(x)$, equation (2.10) can also be written as

$$\left.\frac{dy}{dx}\right|_{x=a} = f'(a),$$

where $\frac{dy}{dx}$ can be taken as an alternative denotation of $f'(x)$, and $\frac{d}{dx}$ is called the **differentiation operator**.

It is easy to prove that a necessary condition for a function $f(x)$ to be differentiable at $x = a$ is that $f(x)$ is continuous at $x = a$.

2.3 Calculation of the Derivative of a Function

The derivative of commonly seen functions are concluded in the following Table 2.1.

With both $f_1(x)$ and $f_2(x)$ differentiable,

$$\begin{aligned}
\frac{d}{dx}\left(af_1(x) + bf_2(x)\right) &= a\frac{d}{dx}f_1(x) + b\frac{d}{dx}f_2(x), \\
\frac{d}{dx}\left(f_1(x)f_2(x)\right) &= f_2(x)\frac{d}{dx}f_1(x) + f_1(x)\frac{d}{dx}f_2(x), \\
\frac{d}{dx}\left(\frac{f_1(x)}{f_2(x)}\right) &= \frac{f_2(x)\frac{d}{dx}f_1(x) - f_1(x)\frac{d}{dx}f_2(x)}{\left(\frac{d}{dx}f_2(x)\right)^2}.
\end{aligned}$$

TABLE 2.1

Derivative of commonly seen functions.

$f(d)$	$f'(x)$	Comments
c	0	
x^n	nx^{n-1}	$n \neq 0$
$\sin(x)$	$\cos(x)$	
$\cos(x)$	$-\sin(x)$	
a^x	$a^x \ln a$	$a > 0$
e^x	e^x	A special case of a^x
$\log_a x$	$\frac{1}{\ln a}\frac{1}{x}$	$a, x > 0$
$\ln x$	$\frac{1}{x}$	A special case of $\log_a x$

If $f = f(x)$ and $g = g(f)$, $\frac{dg}{dx}$ can be calculated using **chain rule for composite function** as

$$\frac{dg}{dx} = \frac{d}{df}g(f)\frac{d}{dx}f(x).$$

For example, if $f = 3x^2 - 2$ and $g = f^2$,

$$\frac{dg}{dx} = \frac{d}{df}g(f)\frac{d}{dx}f(x) = (2f)\,(6x) = 36x^3 - 24x.$$

There are a bunch of theorems not covered with much details in this notebook. For example, the famous *mean value theorem* says if a function $f(x)$ is continuous in $[a, b]$ and has a derivative everywhere in (a, b), there must be such $a < c < b$ that

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

Some of these theorems are widely used in the derivation and proof of other theorems.

3

Integral

CONTENTS

3.1 A Motivating Example 23
3.2 Integral of a Function 28
3.3 Calculation of the Integral of a Function 30

Consider a continuous function $f(x)$ defined in $[a, b]$. We want to find the "antiderivative" of $f(x)$, i.e. find a function $F(x)$ whose derivative is $f(x)$. In Section 3.1, a motivating example is given to illustrate a use case for such "antiderivative". The formal definition of integral is given in Section 3.2, and the calculation of integral for common functions are given in Section 3.3.

3.1 A Motivating Example

Consider the following motivating example where we would like to calculate the area between $y = x^2$ and $y = 0$ for $0 \le x \le 1$.

A Motivating Example

Consider

$$y = x^2, 0 \le x \le 1. \tag{3.1}$$

A plot of (3.1) is given in Fig. 3.1. To obtain the area of the red shape, approximate the red area with the sum of N trapezoids, as shown by the blue dashed curves in Fig. 3.1.

Q1: Calculate the sum of the area of the N trapezoids.

Q2: On top of Q1, let $N \to \infty$ to obtain the area of the red shape.

Q3: Given $0 \le a \le 1$, calculate the area of the trapezoids in between two vertical lines $x = 0$ and $x = a$.

Q4: Given $0 \le a \le 1$, calculate the area of the red in between two vertical lines $x = 0$ and $x = a$.

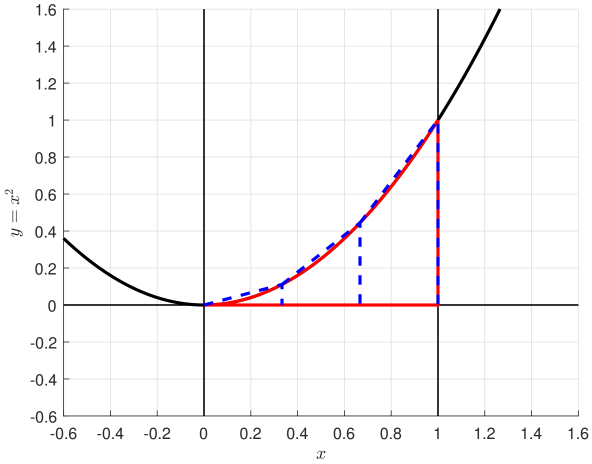

FIGURE 3.1
Plot of (3.1) and $N = 3$ trapezoids.

Notice that in Fig. 3.1, $N = 3$ trapezoids is used in the plot for clearer demonstration. In practice, more trapezoids shall be considered to get a better approximation. We can all agree on that with a larger choice of N, a better approximation can be obtained. When $N \to \infty$, the blue dashed trapezoids approaches the red shape geometrically.

The area of the i-th trapezoid can be calculated as follows. (Let the left side smallest triangle be the first trapezoid, and the right side largest trapezoid be the N-th trapezoid.)

$$\begin{aligned} \Delta x &= \frac{1}{N}, \\ S_i &= \frac{((i-1)\Delta x)^2 + (i\Delta x)^2}{2}\Delta x, \end{aligned}$$

where Δx, $((i-1)\Delta x)^2$ and $(i\Delta x)^2$ are the altitude, shorter base and longer base of the i-th trapezoid, respectively.

Therefore, the total area of the trapezoids is

$$\begin{aligned}
s_T &= \sum_{i=1}^{N} S_i \\
&= \sum_{i=1}^{N} \frac{((i-1)\Delta x)^2 + (i\Delta x)^2}{2}\Delta x \\
&= \sum_{i=1}^{N} \frac{\left((i-1)\frac{1}{N}\right)^2 + \left(i\frac{1}{N}\right)^2}{2}\frac{1}{N} \\
&= \frac{\sum_{i=1}^{N}\left(2i^2 - 2i + 1\right)}{2N^3}. \qquad (3.2)
\end{aligned}$$

Using Table 1.3, equation (3.2) becomes

$$\begin{aligned}
s_T &= \frac{\frac{N(N+1)(2N+1)}{3} - N(N+1) + N}{2N^3} \\
&= \frac{2N^3 + N}{6N^3}. \qquad (3.3)
\end{aligned}$$

Equation (3.3) gives the sum of the area of the trapezoids. For example, substituting $N = 3$ into (3.3) gives $s_T = 0.3519$. This illustrates the case where 3 trapezoids are used to approximate the area of the red as shown in Fig. 3.1, and the approximated area is 0.3519. With larger N, a better estimation can be obtained. For example, $N = 5$ gives $s_T = 0.3400$ and $N = 10$ gives $s_T = 0.3350$ and $N = 20$ gives $s_T = 0.3337$. A plot of using 20 trapezoids to approximate the red area is given in Fig. 3.2. Comparing the two figures 3.1 and 3.2, we can see that $N = 20$ is a fairly good approximation and it gives more accurate result than $N = 3$.

To calculate the precise red area, let N approaches infinity in (3.3) to get

$$\begin{aligned}
S &= \lim_{N\to\infty} \frac{2N^3 + N}{6N^3} \\
&= \lim_{N\to\infty} \frac{2 + \frac{1}{N^2}}{6} \\
&= \frac{1}{3}.
\end{aligned}$$

The calculation of the trapezoids area in between $x = 0$ and $x = a$, $0 \le a \le 1$ is tedious but not theoretically difficult. The basic idea is to discuss a in each single trapezoid individually, and finally use a piece-wise function to represent the area. The larger N is, the more segments it will require in the result.

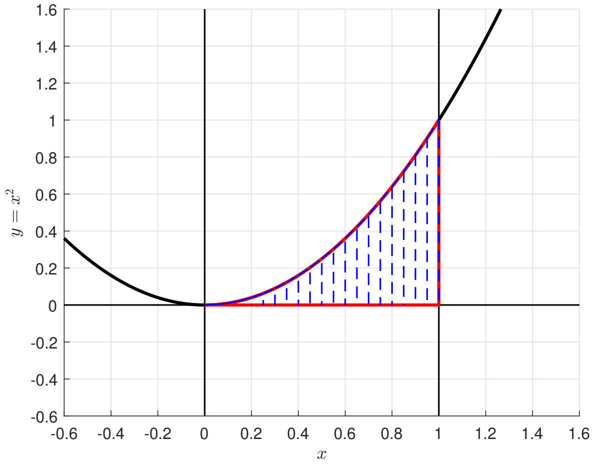

FIGURE 3.2
Use $N = 20$ trapezoids to approximate the red area.

Here we simply use $N = 3$ as an example. The result for any arbitrarily large N can be obtained similarly, just with more calculation and probably even support from a computer.

The upper edge of the trapezoids for $N = 3$ is given by 3.4. It is plotted as the solid blue line in Fig. 3.3.

$$z_3 = \begin{cases} \frac{1}{3}x & 0 \leq x \leq \frac{1}{3} \\ x - \frac{2}{9} & \frac{1}{3} < x \leq \frac{2}{3} \\ \frac{5}{3}x - \frac{2}{3} & \frac{2}{3} < x \leq 1 \end{cases} . \tag{3.4}$$

The area of the trapezoids in between $x = 0$ and $x = a$ for $N = 3$ is given by the red dashed area in Fig. 3.3, where $a = 0.5$ is used in the plot just for demonstration. The area as a function of a is given as

$$s_T^a = \begin{cases} \frac{1}{6}a^2 & 0 \leq a \leq \frac{1}{3} \\ \frac{1}{2}a^2 - \frac{2}{9}a + \frac{1}{27} & \frac{1}{3} < a \leq \frac{2}{3} \\ \frac{5}{6}a^2 - \frac{2}{3}a + \frac{5}{27} & \frac{2}{3} < a \leq 1 \end{cases} . \tag{3.5}$$

The following can be observed from (3.4) and (3.5).

- Equation (3.5) is only an approximation to the area in $x = 0$, $x = a$, $y = 0$ and $y = x^2$. This is because the blue dashed trapezoids are only an approximation to $y = x^2$ given by the black bashed line in Fig. 3.3.

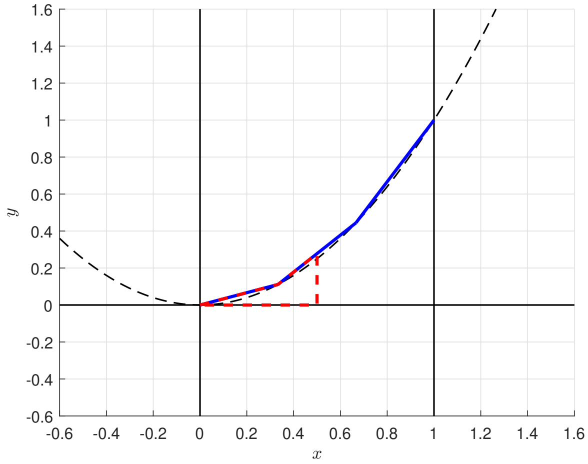

FIGURE 3.3
Calculation of the area of trapezoids. Variables $N = 3$ and $a = 0.5$ are used in the plot for demonstration.

- By increasing N, equation (3.5) gives a better and better approximation. When $N \to \infty$, the accurate area in $x = 0$, $x = a$, $y = 0$ and $y = x^2$ can be obtained.
- VERY IMPORTANT: Equation (3.4) is the derivative of (3.5) for $0 < x < 1$, if replacing the notation "a" in (3.5) with "x".

The last statement above might be difficult to comprehend and explain, but it can be verified rather easily using (3.4) and (3.5).

Notice that in this motivating example, $y = x^2$, $0 < a < 1$ and $N = 3$ are chosen for demonstration purpose without losing generality, thus, this statement shall hold true for any continuous $y = f(x)$ in (3.1) and any arbitrarily large N. We could have chosen $N \to \infty$ for any $y = f(x)$ to calculate the area in $x = a$, $x = b$, $y = 0$ and $y = f(x)$ to get precise (not approximated, since $N \to \infty$) result.

Furthermore, when $N \to \infty$, we know that the trapezoids edge given by (3.4) will perfectly overlap $y = f(x)$. And we know that (3.4) will always be the derivative of the area equation (3.5). Thus, we can derive (3.5) for the $N \to \infty$ case by simply looking for a function $F(x)$ whose derivative is $f(x)$. After that, substituting a and b into $F(b) - F(a)$ gives the precise area surrounded by $x = a$, $x = b$, $y = 0$ and $y = f(x)$.

An intuitive explanation to this statement is given as follows.

The increment of the area function $F(x)$ from any value $x = a$ to $x = a + \Delta x$ is by definition the area in between $x = a$, $x = a + \Delta x$, $y = 0$ and $y = f(x)$. Since Δx is very small (in fact, when $N \to \infty$, $\Delta x \to 0$),

$f(a+\Delta x) \to f(a)$ and this small area is $f(a) \times \Delta x$. On the other hand, the derivative of $F(x)$ at $x = a$ is defined as this increment $f(a) \times \Delta x$ divided by Δx. Therefore, the derivative of $F(x)$ at $x = a$ is $f(a)$. Alternatively, we can say *the area function $F(x)$ for $y = f(x)$ is the "antiderivative" of $f(x)$*, and in the next section we will meet its official name, the "integral".

3.2 Integral of a Function

In this section, the definitions of *definite integral* and *indefinite integral* are given. As a first step, Riemann integral is introduced. Then, Riemann integral is "translated" into the definition of definite integral.

Definition of Riemann integral:

Consider a function $f(x)$ defined on interval $[a, b]$. Let $[a, b]$ be split into N consecutive segments whose length are given by $\lambda_1, ..., \lambda_N$, with the longest segment's length being $\lambda = \max\{\lambda_1, ..., \lambda_N\}$. Let x_i be a sample randomly taken inside the i-th segment.

If for any arbitrarily small ε, there is always such δ that as long as $\lambda < \delta$,

$$\left| \sum_i f(x_i)\lambda_i - S \right| < \varepsilon,$$

for a constant S, then S is called the Riemann integral of function $f(x)$ on interval $[a, b]$.

The segment splitting and sampling used in Riemann integral is more general than what Section 3.1 has been using. In Section 3.1, we have been assuming even splitting segment $\lambda_i = \lambda_j$ for any two segments, and also determinant sampling in the segment $x_i = \underset{x}{\arg} f(x) = \frac{1}{2}\left(f(x_l) + f(x_r)\right)$ where x_l, x_r are the left and right boundary of the segment respectively. By assuming continuous function $f(x)$ in Section 3.1, it is guaranteed that such x_i exists. The approach used in Section 3.1 is only a special case of Riemann integral, but it would work just fine for most of the cases.

A more intuitive definition of definite integral is given below. The idea is inherited from the motivating example in Section 3.1. It is not as "strong" as the Riemann's definition, but should be sufficient for most of the use cases.

Definition of definite and indefinite integrals in an intuitive manner:

Consider a continuous function $f(x)$ defined on interval $[a, b]$. Let $[a, b]$ be split into N consecutive segments with equal length Δx. Let x_i be a sample in the i-th segment. The *definite integral* of $f(x)$ on interval $[a, b]$ is given by the following equation

$$S = \lim_{\Delta x \to 0} \sum_i f(x_i)\Delta x, \tag{3.6}$$

if the right side limit exists.

In such case, we can rewrite (3.6) with the following denotations.

$$S = \int_a^b f(x)dx. \tag{3.7}$$

where a and b are called the lower and upper bound of the integral, respectively. The "dx" in (3.7) can be taken as $\Delta x \to 0$.

Equation (3.7) can be solved by finding such $F(x)$ that $\frac{d}{dx}F(x) = f(x)$, and

$$S = \int_a^b f(x)dx = F(b) - F(a). \tag{3.8}$$

Function $F(x)$ is called the *indefinite integral* of function $f(x)$, and it is denoted by

$$F(x) = \int f(x)dx,$$

which does not come with the lower and upper bound, and it is often not unique.

The integral may not exist for some $f(x)$, or at the very least it is impossible to derive an analytical equation $F(x)$ associated with that $f(x)$. It is often easier to derive the derivative of a function, i.e. from $F(x)$ to $f(x)$, than the other way around.

Notice that in the definitions of definite and indefinite integral, continuous $f(x)$ is assumed. For those piece-wise functions that is not continuous at certain values in its range, particular caution is required, for example, to analyze it piece-by-piece considering boundary conditions.

TABLE 3.1
Indefinite integral of commonly seen functions.

$f(x)$	$F(x) = \int f(x)dx$	Comments
a	$ax + c$	
x^n	$\frac{1}{n+1}x^{n+1} + \mathrm{c}$	$n \neq -1$
x^{-1}	$\ln\lvert x\rvert + c$	$x \neq 0$
$\frac{1}{ax+b}$	$\frac{1}{a}\ln\lvert ax+b\rvert + c$	$a \neq 0$, $ax + b \neq 0$
$\sin(x)$	$-\cos(x) + c$	
$\cos(x)$	$\sin(x) + c$	
e^x	$e^x + c$	
$\ln x$	$x\ln x - x + c$	
$\frac{1}{\sqrt{1-x^2}}$	$\frac{1}{\sin(x)} + c$	$\lvert x\rvert < 1$
$\frac{1}{1+x^2}$	$\frac{\cos(x)}{\sin(x)} + c$	

In the case of Riemann integral, the assumption is more relaxed as continuity of $f(x)$ is not required. However, it is still possible that for some $f(x)$, Riemann integral does not exist. For example, Dirichlet function,

$$D(x) = \begin{cases} 1 & x \in \mathbb{Q} \\ 0 & \text{otherwise} \end{cases}, \tag{3.9}$$

is not continuous or differentiable at any x, and it is not Riemann integrable on any interval.

3.3 Calculation of the Integral of a Function

Some commonly seen indefinite integral is given in Table 3.1. When calculating indefinite integral, there is always an arbitrary constant c in the result, as the constant in $F(x)$ does not affect the derivative $f(x)$.

If a function is similar to the functions presented in Table 3.1, its indefinite integral might be achievable.

The following rules are commonly used when calculating the integral.

$$\begin{aligned} \int f(x) + g(x) &= \int f(x)dx + \int g(x)dx + c, \\ \int af(x)dx &= a\int f(x)dx. \end{aligned}$$

The integral can be calculated by parts as follows.

$$\int u(x)v'(x)dx = uv - \int u'(x)v(x)dx,$$

where $u'(x)$, $v'(x)$ are the derivative of $u(x)$ and $v(x)$ respectively. For example, to solve the integral of $f(x) = x\text{sin}(x)$, let $u(x) = x$, $v'(x) = \text{sin}(x)$. From Table 3.1, we know that $u'(x) = 1$, and $v(x) = -\text{cos}(x) + c$.

$$\begin{aligned}\int f(x)dx &= \int u(x)v'(x)dx \\ &= u(x)v(x) - \int u'(x)v(x)dx \\ &= x\left(-\text{cos}(x) + c_1\right) - \int \left(-\text{cos}(x) + c_1\right)dx \\ &= -x\text{cos}(x) + \int \text{cos}(x)dx \\ &= -x\text{cos}(x) + \text{sin}(x) + c.\end{aligned}$$

Inspired by the chain rule of calculation of derivative, the integral can be calculated by substitution as follows.

$$\int f\left(g(x)\right)dg(x) = \int f\left(g(x)\right)g'(x)dx = F\left(g(x)\right) + c,$$

where $F(x) = \int f(x)dx$.

In general, the definite integral of a continuous function can be calculated by substituting the upper and lower bound into the indefinite integral and do the following subtraction

$$\int_a^b f(x)dx = -\int_b^a f(x)dx = F(b) - F(a).$$

Notice that sometimes $F(b) - F(a)$ is also denoted by $F(x)|_a^b$.

As illustrated in the motivating example in Section 3.1, for $a < b$, $\int_a^b f(x)dx$ can be interpreted as the accumulated area between $y = f(x)$ and $y = 0$, within boundary $x = a$ and $x = b$.

When $f(x) \geq 0$, $\int_a^b f(x)dx \geq 0$ is the area surrounded by $y = f(x)$, $y = 0$, $x = a$ and $x = b$. When $f(x) \leq 0$, $\int_a^b f(x)dx \leq 0$ is the negative of the area. Otherwise, $\int_a^b f(x)dx$ is the difference of area above and below $y = 0$ as shown in Fig. 3.4.

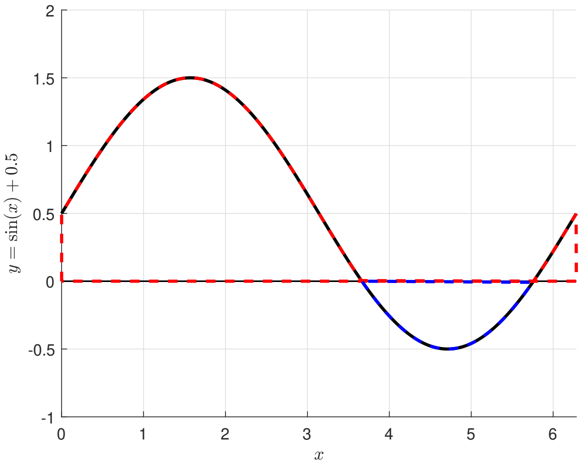

FIGURE 3.4
Plot of $f(x) = \sin(x) + 0.5$ from 0 to 2π. The area above $y = 0$ is surrounded by the red dashed line, and the area below $y = 0$ by the blue dashed line. The definite integral $\int_0^{2\pi} f(x)dx$ in this case is the red area subtracting the blue area.

4

Applications

CONTENTS

4.1 Newton's Method 33
4.2 Taylor Series 34

This chapter introduces some interesting and widely known use cases of derivatives and integrals.

Section 4.1 introduces Newton's method, a widely used numerical method to solve equation $f(x) = 0$ for some continuous function $f(x)$.

Section 4.2 introduces Taylor serious, a widely used numerical method to approximate the value of $f(x)$ near a specific x_0.

4.1 Newton's Method

Consider solving an equation $f(x) = 0$ for continuous function $f(x)$. Sometimes it is possible to construct a function $g(x)$, such that $x = x + g(x)$ has the same solution with $f(x) = 0$. Equation $f(x) = 0$ might then be solved recursively using $x^{k+1} = x^k + g(x^k)$ where k is the recursive index. In this notebook, we are not going to discuss how $g(x)$ can be constructed and what limitations of this method may have.

In Newton's method, $g(x)$ is constructed as $g(x) = -\frac{f(x)}{f'(x)}$. For Newton's method to converge to the correct solution, there is some restrictions to $f(x)$ and also the choice of x^0 as the initial guess. The detailed discussion to these restrictions are not covered in this notebook.

The basic procedures for Newton's method are given below.

Step 1: Determine a feasible range $[a, b]$ **where the solution to** $f(x) = 0$ **must lie inside.** This can be done by having a rough guess of an range $[a, b]$ near the solution, and make sure $f(a)f(b) < 0$. It can be proved that for a continuous function $f(x)$, $f(a)f(b) < 0$ guarantees a solution in $[a, b]$.

Step 2: Have a initial guess $x^0 \in [a, b]$**. Initialize** $k = 0$**.**

Step 3: Calculate $f'(x^k)$, then calculate x^{k+1} as follows.

$$x^{k+1} = x^k - \frac{f(x^k)}{f'(x^k)}.$$

Step 3 is iterated to for recursive calculation of x^k. The iterations stops when either of the following happens: (a) $|f(x^k)|< \varepsilon$; or (b) $|x^{k+1} - x^k|< \varepsilon$, where ε is a pre-defined threshold parameter. The finally calculated x^k is the numerical solution of the original equation $f(x) = 0$ using Newton's method.

4.2 Taylor Series

Consider a continuous and $(n+1)$-th order differentiable function $f(x)$ defined on interval $[a, b]$. Taylor series introduces a way to approach a function $f(x)$ at any $x \in [a, b]$ using a sum of sequence consisting $f(x)$ and its derivatives at a particular $x_0 \in [a, b]$.

Taylor series claims that such $f(x)$ can be expressed by the following equation

$$f(x) = P(x, x_0), +R(x, x_0), \tag{4.1}$$

where

$$\begin{aligned} P(x, x_0) &= \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k, \\ R(x, x_0) &= \frac{f^{(n+1)}\left(x_0 + \theta(x - x_0)\right)}{(n+1)!}(x - x_0)^{n+1}, \end{aligned} \tag{4.2}$$

with $\theta \in (0, 1)$, and $f^{(n)}(x)$ the n-th order derivative of $f(x)$, i.e. "the derivative of derivative of ... derivative of $f(x)$", where there are n "derivative" in the sentence.

Equation (4.1) uses $P(x, x_0)$ to approximate $f(x)$. It can be seen from (4.2) that with $n \to \infty$ or $x \to x_0$, the remainder $R(x, x_0) = f(x) - P(x, x_0)$ approaches zero (notice that $f^{(n+1)}$ is assumed bounded). This implies that the approximation performs better when higher order Taylor series is used, or when the target point x is close to the evaluated point x_0.

The following Fig. 4.1 gives an example of using Taylor series to approximate function $y = 2^x$ at $x_0 = 0$. In this example, it can be seen that the approximation gets better when higher order Taylor series is used.

For polynomial functions $f(x)$, $f^{(n)}(x) = 0$ for large enough n, depending on the degree of the polynomial. For a N-th-order polynomial, $f^{(n)}(x) = 0$ for $n > N$, thus its n-th-order (or higher) Taylor series $P(x, x_0) = f(x)$ for any choice of x_0.

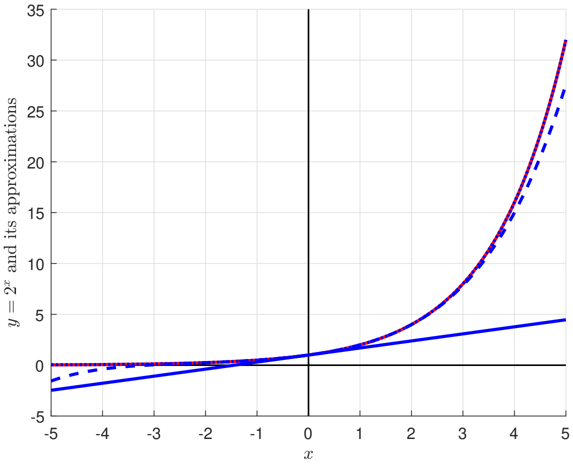

FIGURE 4.1
Plot of function $y = 2^x$ in red solid line, and its approximations using first-order, 5th-order and 10th-order Taylor series in blue solid line, blue dashed line and blue dot line respectively.

Part II

Multivariable Function

5

Multivariable Function

CONTENTS

5.1 Brief Introduction to Vector and Matrix 39
 5.1.1 Basic Concepts 40
 5.1.2 Matrix Multiplication 41
 5.1.3 Block Matrix 42
 5.1.4 Identity Matrix and Square Matrix Inverse 43
5.2 Multivariable Function 43

This chapter introduces functions with multiple inputs and/or outputs. Usually, these inputs and outputs are put into vectors for computation and presentation convenience. Section 5.1 gives a very brief introduction to the basics of vector and matrix operations. Section 5.2 introduces the concept of multivariable functions, including the multiple input function and the vector function.

From calculus perspective, this chapter clears the preliminary knowledge required for later Chapters 6 and 7.

5.1 Brief Introduction to Vector and Matrix

Detailed introduction to vector and matrix can be found in any *linear algebra* textbook, where there are the geometric interpretation of vector, then linear equation represented by product of matrix and vector, then column and null space of matrix, then rank of matrix and determinant of square matrix, then inverse of matrix, then linear transformation of matrix, then eigenvalue of matrix, then norm of vector and matrix, then algebraic Riccati equation, and so on. This list can go almost eternally. To make things even more complicated, depending on the application, the vector and matrix may have different physical meanings. For example, the speed of motion and the bus voltage phasors of a microgrid can all be represented by vectors, but might be completely two different things.

In the context of this notebook, however, most of the above are out of the scope. We will take vector and matrix as a way of organizing scalar data. Basically, a vector is a 1-D chain of scalar variables, and a matrix is a 2-D rectangular mesh of scalar variables. In many cases such as calculating matrix product, a vector can be taken as a special case of matrix, and a scalar is a special case of a vector. Of course, the very basics such as product of matrices are still required.

The preliminary vector and matrix knowledge used in this notebook is summarized in the rest of this section as follows.

5.1.1 Basic Concepts

A vector x (sometimes denoted as bold text $\mathbf{x}$ in textbooks) is a finite sequence of scalar variables organized in a 1-D chain. The length of the vector, or the dimension of the vector, is the number of elements in the vector. For vector x with n elements, the elements are denoted by $x_1, x_2, \cdots, x_n$, and x_i is called the i-th element of the vector.

Most of the vectors in this notebook are by default *column vectors*, which means that the n elements are put into n-row-1-column, as follows. In this case, we say *"x is a n dimensional column vector"* or *"x is a $n \times 1$ vector"*.

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}. \tag{5.1}$$

A row vector, on the other hand, puts the n elements into 1-row-n column. A row vector is like a column vector flipped diagonally, as shown below in (5.2). We call the flipping operation *"transpose"*, denoted by $(\cdot)^T$ (used in this textbook) or $(\cdot)'$.

$$x^T = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}, \tag{5.2}$$

where x^T is a row vector and it is a transpose of the column vector x previously given in (5.1). The transpose of a column vector is a row vector, and vise versa. The transpose of a scalar is itself.

A matrix A (sometimes denoted as bold $\mathbf{A}$ in textbooks) is a finite number of scalar variables organized in 2-D mesh, with each scalar taking a particular position. The number of rows and columns are the dimension of the matrix. For example, if A has m rows and n columns, we say *"A has a dimension of $m \times n$"* or *"A is a $m \times n$ matrix"*, shown as follows.

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \dots & a_{m,n} \end{bmatrix}, \tag{5.3}$$

where $a_{i,j}$ is the i-th row, j-th column element of A.

The transpose operation is also defined on matrix. By applying transpose on A in (5.3), a $n \times m$ matrix A^T can be obtained, where the i-th row, j-th column element in the transpose matrix A^T is the j-th row, i-th column element in the original matrix A.

The vectors or matrices with the same dimension can be added together by adding each associated pair of elements together.

5.1.2 Matrix Multiplication

The product of two matrices is defined as follows.

Matrix Multiplication:

Consider matrices A and B. As a prerequisite of calculating AB, the number of column in the first matrix A must equal to the number of row in the second matrix B.

Let A and B be $m \times p$ matrix and $p \times n$ matrix respectively. The matrix product $C = AB$ is a $m \times n$ matrix with each element calculated by

$$\begin{aligned} c_{i,j} &= a_{i,1}b_{1,j} + a_{i,2}b_{2,j} + \ldots + a_{i,p}b_{p,j} \\ &= \sum_{k=1}^{p} a_{i,k}b_{k,j}, \end{aligned}$$

for $i = 1, \ldots, m$ and $j = 1, \ldots, n$.

It is clearly from the definition that AB does not equal to BA. In the example above, if $m \neq n$, BA does not exist in the first place.

It can be proved by definition that if $C = AB$, $C^T = (AB)^T = B^T A^T$.

In terms of matrix multiplication, a vector is just a special case of matrix. Therefore, the product of a matrix with a vector $y = Ax$ where A is a $m \times n$

matrix and x a $n \times 1$ matrix, is a $m \times 1$ vector given by

$$\begin{aligned} y &= Ax \\ &= \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \dots & a_{m,n} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \\ &= \begin{bmatrix} \sum_{k=1}^{n} a_{1,k}x_k \\ \sum_{k=1}^{n} a_{2,k}x_k \\ \vdots \\ \sum_{k=1}^{n} a_{m,k}x_k \end{bmatrix}. \end{aligned}$$

The product of row vector u^T and column v, both n dimensional vectors, is

$$\begin{aligned} u^T v &= \begin{bmatrix} u_1 & u_2 & \dots & u_n \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \\ &= \sum_{k=1}^{n} u_k v_k \end{aligned}$$

which is a scalar. As a special case, $x^T x = \sum_{k=1}^{n} x_i^2$ for a n dimensional vector x.

5.1.3 Block Matrix

For the convenience of calculation and interpretation, sometimes a large dimension matrix is split into a combination of smaller dimension sub matrices.

For example, consider matrix A with dimension $m \times p$ and matrix B with dimension $p \times n$. Matrix A can be split into two sub matrix A_1 and A_2, where A_1 consists of the first m_1 rows of A thus $m_1 \times p$ dimension, and A_2 consists of the rest $m_2 = m - m_1$ rows of A thus $m_2 \times p$ dimension, i.e.

$$A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}.$$

The calculation of $C = AB$ can be done as follows

$$C = AB = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} B = \begin{bmatrix} A_1 B \\ A_2 B \end{bmatrix}.$$

Similarly, if split matrix B into two sub matrices, with B_1 the first n_1 columns and B_2 the rest $n_2 = n - n_1$ columns of B, then

$$C = AB = A \begin{bmatrix} B_1 & B_2 \end{bmatrix} = \begin{bmatrix} AB_1 & AB_2 \end{bmatrix}$$

Furthermore, splitting both A and B simultaneously gives

$$C = AB = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \begin{bmatrix} B_1 & B_2 \end{bmatrix} = \begin{bmatrix} A_1B_1 & A_1B_2 \\ A_2B_1 & A_2B_2 \end{bmatrix} \tag{5.4}$$

Equation (5.4) sometimes helps to speed up the matrix product as it allows to split the calculation into independent pieces. But more importantly, equation (5.4) gives a lot of insights into matrix operations and linear transformation. The details are not covered in this notebook.

5.1.4 Identity Matrix and Square Matrix Inverse

A matrix is called a *square matrix* if it has the same number of rows and columns. For example, if matrix A has a dimension of $n \times n$, then it is a square matrix of dimensional n. The elements with the same row and column index, i.e. $a_{i,i}, i = 1, \ldots, n$, are called the diagonal elements.

The *identity matrix*, denoted by I, is a special type of square matrix as given in (5.5). Its diagonal elements are 1, with the rest elements 0.

$$I = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix} \tag{5.5}$$

A n dimensional identity matrix is denoted by I_n. The multiplication of the identity matrix with any matrix from either left or right side does not change the matrix, i.e. for any matrix A with dimension $m \times n$, $I_mA = AI_n = A$.

A square matrix may have an associated inverse matrix. The definition of inverse matrix is given as follows.

Definition of inverse of a matrix:

Consider a square matrix A with dimension $n \times n$. Matrix A is called *invertible* if such matrix A^{-1} with dimension $n \times n$ exists that

$$AA^{-1} = A^{-1}A = I_n,$$

and A^{-1} is called the *inverse* of matrix A.

Notice that matrix A^{-1} may not exist for some A, depending on the determinant of A. Details can be found in linear algebra textbooks and are not given in this notebook.

5.2 Multivariable Function

In Chapters 2 and 3, we have been considering single-input-single-output functions only, i.e. for function $y = f(x)$, we have been discussing only the cases where y and x are scalars so far.

In many other cases, however, a function may have multiple input and/or output variables. For example, consider the following function used to calculate the mechanical energy of an object

$$e = f(v, h) = \frac{1}{2}mv^2 + mgh.$$

where m, v and h are the mass, motion speed and height (related to ground) of the object, and g is the free-fall acceleration, $g = 9.8m/s^2$ on the earth. This is a typical multivariable function, where the function $z = f(x, y)$ depends on multiple independent variables x and y.

When there are multiple outputs for a function, the outputs are often out into a column vector and the function is called a *vector function*. Two examples of vector functions are given below.

$$f(x) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 3 & -2 \end{bmatrix} x, \tag{5.6}$$

$$g(x) = \begin{bmatrix} g_1(x) \\ g_2(x) \end{bmatrix} = \begin{bmatrix} x_1^2 + x_2^2 \\ 0.5x_1 + e^{x_2} \end{bmatrix}, \tag{5.7}$$

where $f(x)$ and $g(x)$ are both vectors with dimension 3 × 1 and 2 × 1 respectively. Equation (5.6) is a linear function with 2 inputs $x = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^T$ and 3 outputs $y = \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix}^T$, while (5.7) is a nonlinear function with 2 inputs $x = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^T$ and 2 outputs $g = \begin{bmatrix} g_1 & g_2 \end{bmatrix}^T$.

6

Partial derivative

CONTENTS

6.1 A Motivating Example 45
6.2 Partial Derivative 48
6.3 Gradient 51
  6.3.1 Motivating Example 51
  6.3.2 Gradient of a Multivariable Function 56
6.4 Jacobian Matrix 58

Partial derivative studies the effect of a small deviation of one particular independent variable on the multivariable function. It is similar with the normal derivative in many ways, but also has its unique characteristics.

In Section 6.1, a motivating example is given to illustrate the motivation of introducing partial derivative. In Section 6.2, the definition of partial derivative is given. In Sections 6.3 and 6.4, two very important and commonly used tool derived from partial derivative, namely gradient and Jacobian matrix, are introduced respectively.

6.1 A Motivating Example

Consider the following motivating example where $y = f(x_1, x_2)$ is a multivariable function with 2 inputs.

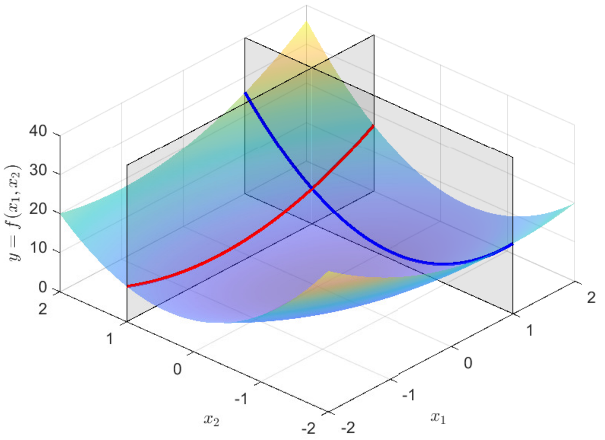

FIGURE 6.1
Plot of function $y = f(x_1, x_2)$ in 3-D.

A Motivating Example

Consider

$$y = 2x_1^2 + x_2^2 + 2x_1x_2. \tag{6.1}$$

Q1: Let $x_2 = 1$ be a constant. Derive y as a function of x_1, and calculate its derivative with respect to x_1. Similarly, let $x_1 = 1$ be a constant and derive y as a function of x_2, and calculate its derivative with respect to x_2.

Q2: At $(x_1, x_2) = (1, 1)$, consider small vibrations Δx_1 and Δx_2. Approximate Δy as a function of Δx_1 and Δx_2 using differentiation.

Q3: Find such x_1 and x_2 that y is minimized.

Equation (6.1) can be plot in 3-D as Fig. 6.1.

Let $x_2 = 1$ be constant. Equation (6.1) becomes

$$y = f(x_1, 1) = 2x_1^2 + 2x_1 + 1,$$

which is given by the intersection of curved surface (equation (6.1)) and the vertical flat ($x_2 = 1$) given by the red solid line in Fig. 6.1. Its derivative with respect to x_1 can be easily obtained as

$$\frac{d}{dx_1} f(x_1, 1) = 4x_1 + 2. \tag{6.2}$$

Similarly, at $x_1 = 1$, (6.1) becomes

$$y = f(1, x_2) = x_2^2 + 2x_2 + 2,$$

which is given by the blue solid line in Fig. 6.1. Its derivative with respect to x_2 is

$$\frac{d}{dx_2}f(1,x_2) \quad = \quad 2x_2+2. \tag{6.3}$$

Consider small vibrations Δx_1 and Δx_2. Firstly, let x_2 remain constant and only apply Δx_1 on $y=f(1,1)$. In this case,

$$\Delta y = f(1+\Delta x_1, 1) - f(1,1) \approx \frac{d}{dx_1}f(x_1,1)\Delta x_1, \tag{6.4}$$

where $\frac{d}{dx_1}f(x_1,1)$ is given in (6.2).

On top of (6.4), consider vibration Δx_2. The derivative of function $f(1+\Delta x_1, x_2)$ with respect to x_2 depends on Δx_1 and it is generally unknown without specifying Δx_1. However, since (6.1) is continuous and "smooth" (Notice that "smooth" has specific definition in mathematics. Here, just take its literal meaning: it is perfectly polished without any edges or spikes.), we can intuitively understand that when Δx_1 is small, $\frac{d}{dx_2}f(1+\Delta x_1, x_2) \approx \frac{d}{dx_2}f(1,x_2)$, and $\frac{d}{dx_2}f(1+\Delta x_1,x_2) \to \frac{d}{dx_2}f(1,x_2)$ as $\Delta x_1 \to 0$. Thus, we have

$$\begin{aligned}\Delta y &= (f(1+\Delta x_1, 1+\Delta x_2) - f(1+\Delta x_1, 1)) + (f(1+\Delta x_1,1) - f(1,1)) \\ &\approx \frac{d}{dx_2}f(1,x_2)\Delta x_2 + \frac{d}{dx_1}f(x_1,1)\Delta x_1. \end{aligned} \tag{6.5}$$

In (6.5), consider $\Delta x_1 \to 0$ and $\Delta x_2 \to 0$. Denote such Δx_1 and Δx_2 as dx_1 and dx_2 respectively, and the associated Δy as dy. Here "$d(\cdot)$" represents the *infinitesimal change*. Equation (6.5) then becomes

$$dy \quad = \quad \frac{d}{dx_2}f(1,x_2)dx_2 + \frac{d}{dx_1}f(x_1,1)dx_1. \tag{6.6}$$

Equation (6.6) can be extended to more general cases where $(x_1,x_2) = (1,1)$ is not specified. The terms $\frac{d}{dx_1}f(x_1,1)$ and $\frac{d}{dx_2}f(1,x_2)$ needs to be changed accordingly. For example, $\frac{d}{dx_1}f(x_1,1)$ given by (6.2) shall be changed to

$$\left.\frac{d}{dx_1}f(x_1,x_2)\right|_{x_2 \text{ is constant}} \quad = \quad 4x_1+2x_2$$

with x_2 being any arbitrary constant.

The denotation $\left.\frac{d}{dx_1}f(x_1,x_2)\right|_{x_2 \text{ is constant}}$ can sometimes become ambiguous if not handled carefully. One of the reasons could be that "$d(\cdot)$" is used as infinitesimal change as shown before. In this case both dx_1 and dx_2 contribute to $df(x_1, x_2)$, but we would want $\frac{d}{dx_1}f(x_1, x_2)$ here to reflect the deviation of $f(x_1, x_2)$ caused by dx_1 only, and it is not convenient to list down all the other variables and put "is/are constant" everywhere in the equation. Therefore, instead of saying $\left.\frac{d}{dx_1}f(x_1, x_2)\right|_{x_2 \text{ is constant}}$, we denote

$$\frac{\partial}{\partial x_1}f(x_1, x_2) = 4x_1 + 2x_2.$$

The operator $\frac{\partial}{\partial x}$ is called *partial derivative* (with respect to x), often followed by a multivariable function $f(x, y)$ where x is one of its inputs. It is similar to the derivative, but emphasizing that the interested function is multivariable, and the derivative of this function with respect to ONLY ONE variable is being studied (and the rest variables assumed constant). Since $\partial f(x)$ is rarely or never used as the infinitesimal of $f(x)$, it will hopefully be less ambiguous.

The formal definition of partial derivative is given in next Section 6.2.

Equation (6.6) for general (x_1, x_2), therefore, becomes

$$dy = \frac{\partial}{\partial x_2}f(x_1, x_2)dx_2 + \frac{\partial}{\partial x_1}f(x_1, x_2)dx_1. \tag{6.7}$$

The minimum $y = f(x_1, x_2)$ and its associated x_1 and x_2 can be found by rewriting (6.1) as

$$y = x_1^2 + (x_1 + x_2)^2.$$

Therefore, the minimum y is $y = 0$, at $x_1 = 0$ and $x_1 + x_2 = 0$, i.e., $x_1 = x_2 = 0$.

This can also be solved using (6.7). At the minimum point of y, $\frac{\partial}{\partial x_2}f(x_1, x_2)$ and $\frac{\partial}{\partial x_1}f(x_1, x_2)$ must be both zero. Otherwise, it is always possible to add/subtract a small Δx_1 or Δx_2 to further decrease y. Thus, from (6.1)

$$\begin{aligned}\frac{\partial}{\partial x_1}f(x_1, x_2) &= 4x_1 + 2x_2 = 0 \\ \frac{\partial}{\partial x_2}f(x_1, x_2) &= 2x_1 + 2x_1 = 0\end{aligned}$$

which yields $x_1 = x_2 = 0$.

6.2 Partial Derivative

In the case of single input scalar function $f(x), x \in \mathbb{R}$, there is no point to define partial and total derivative, as the derivative with respect to that single variable by itself is the total derivative.

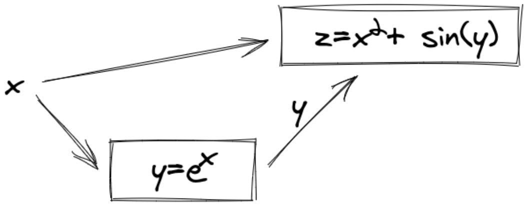

FIGURE 6.2
The calculation of z using x and y.

In the case of multivariable function with multiple inputs, $f(x), x \in \mathbb{R}^{R\times 1}$, the definition partial derivative is given below.

Definition of Partial Integral:

Consider function $y = f(x_1, ..., x_n)$ where y is the scalar output and $x = [x_1, ..., x_n]^T$ is a $n \times 1$ vector input. The partial derivative of $y = f(x)$ with respect to x_i is given by

$$\frac{\partial}{\partial x_i} f(x) = \lim_{\Delta x_i \to 0} \frac{f(x_1, ..., x_i + \Delta x_i, ..., x_n) - f(x_1, ..., x_n)}{\Delta x_i}, \tag{6.8}$$

with $x_1, ..., x_{i-1}, x_{i+1}, ..., x_n$ remaining constant, and only x_i is allowed to vary.

Here is another example to help better understand partial derivative. Consider $z = f(x, y)$ a multivariable function of both inputs x, y as follows

$$z = f(x, y) = x^2 + \sin(y), \tag{6.9}$$

where $y = g(x)$ is a single input function of x as follows

$$y = g(x) = e^x. \tag{6.10}$$

Apparently, the value of z ultimately depends on x alone, but in the intermediate calculation, z depends on both x and y. The calculation flow is visualized in Fig. 6.2.

In this case, the total derivative $\frac{d}{dx} f(x, y)$ can be expressed as follows

$$\frac{d}{dx} f(x, y) \quad = \quad \frac{\partial}{\partial x} f(x, y) + \frac{\partial}{\partial y} f(x, y) \frac{d}{dx} y. \tag{6.11}$$

where in this example,

$$\begin{aligned}\frac{\partial}{\partial x}f(x,y) &= 2x \\ \frac{\partial}{\partial y}f(x,y) &= \cos(y) \\ \frac{d}{dx}y &= e^x\end{aligned}$$

and finally

$$\frac{d}{dx}f(x,y) = 2x + \cos(e^x)e^x$$

Sometimes (6.11) is rewritten as follows

$$\begin{aligned}df(x,y) &= \frac{\partial}{\partial x}f(x,y)dx + \frac{\partial}{\partial y}f(x,y)dy \\ dy &= \left(\frac{d}{dx}y\right)dx\end{aligned}$$

where $d(\cdot)$ represents the infinitesimal change of a variable.

Equation (6.11) serves as a good example to show the relationship and difference between the total derivative $\frac{d}{dx}f(x,y)$ and the partial derivative $\frac{\partial}{\partial x}f(x,y)$. When calculating total derivative, all variables that would affect the value of the function must be taken into consideration, while when calculating partial derivative, only one input variable is studied, with the rest variables remaining constant.

For a function with 1 output and n inputs, sometimes it is convenient to put the partial derivative to each input in a vector. For example, for $y = f(x)$ where $x = [x_1, ..., x_n]^T$, denote

$$\frac{\partial}{\partial x}f(x) = \left[\begin{array}{ccc} \frac{\partial}{\partial x_1}f(x) & \ldots & \frac{\partial}{\partial x_n}f(x) \end{array}\right] \tag{6.12}$$

as the *scalar-by-vector derivative*. Notice that (6.12) is usually given as a row vector. In some research papers the result is given as a column vector, i.e. the transpose of (6.12).

For a function with m outputs and 1 input $y = f(x)$ where $y = [y_1, ..., y_m]^T = [f_1(x), ..., f_m(x)]^T$, its *vector-by-scalar derivative* is given by

$$\frac{\partial}{\partial x}f(x) = \left[\begin{array}{c} \frac{\partial}{\partial x}f_1(x) \\ \vdots \\ \frac{\partial}{\partial x}f_m(x) \end{array}\right]. \tag{6.13}$$

And finally for a function with m outputs and n inputs $y = f(x)$ where $y = [y_1, ..., y_m]^T = [f_1(x), ..., f_m(x)]^T$ and $x = [x_1, ..., x_n]^T$, the *vector-by-vector* derivative is given by

$$\frac{\partial}{\partial x} f(x) = \begin{bmatrix} \frac{\partial}{\partial x_1} f_1(x) & \dots & \frac{\partial}{\partial x_n} f_1(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} f_m(x) & \dots & \frac{\partial}{\partial x_n} f_m(x) \end{bmatrix}. \tag{6.14}$$

Partial derivative and the above equations (6.12), (6.13) and (6.14) have many applications. Two of those most popular use cases are *gradient* and *Jacobian matrix*. Since they are so important and widely used, it is worth especially introducing them in specific Sections 6.3 and 6.4.

6.3 Gradient

For a scalar-valued multivariable function $y = f(x)$ where y is a scalar and x is a vector $x = \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix}^T \in \mathbb{R}^n$, the gradient studies the "direction" of Δx in $\mathbb{R}^n$ space that causes y to increase/decrease the fastest.

The gradient of such function $f(x)$ is a vector function of x, denoted by $\nabla f(x)$. Notice that $\nabla f(x) \in \mathbb{R}^n$ is a vector in the same space with x, since it indicates a "direction" of x.

Section 6.3.1 gives a motivating example of gradient, and Section 6.3.2 gives its formal definition.

6.3.1 Motivating Example

The following motivating example helps to illustrate the calculation and use case of gradient.

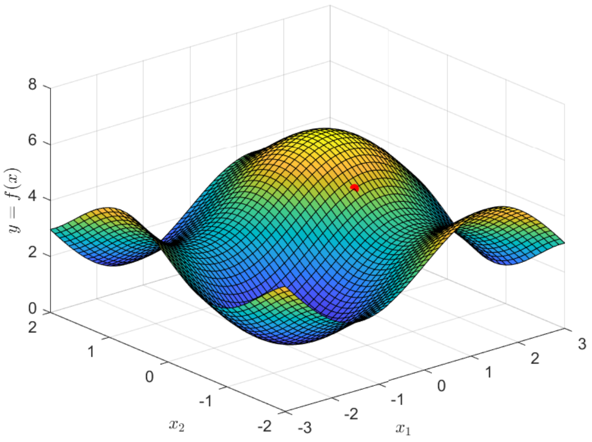

FIGURE 6.3
Plot of $y = f(x_1, x_2)$ in 3-D.

A Motivating Example

Consider the following function $y = f(x)$ where $x = [x_1, x_2]^T$ is a vector and

$$y = f(x) = 2\sin(x_1) + \sin\left(\frac{x_1}{2} + \pi\right) + \sin(2x_2) \tag{6.15}$$

where $x_1 \in [-3, 3]$ and $x_2 \in [-2, 2]$. The 3-D plot and the contour line of the function are given in Figs 6.3 and 6.4 respectively.

Consider initial point $x^0 = [x_1^0, x_2^0]^T = [1, 0]^T$. Let x deviate a little bit from x^0 to get $x^1 = [x_1^1, x_2^1]^T = [x_1^0 + \Delta x_1^0, x_2^0 + \Delta x_2^0]^T$. The objective is to find such $\Delta x^0 = [\Delta x_1^0, \Delta x_2^0]^T$ to hopefully get $f(x^1)$ as large as possible.

Notice that Δx^0 can be interpreted as a vector that points to the "direction" of x where y increases the fastest. An intuitive way is to find the tangent plane to the surface given by (6.15) at $x^0 = [1, 0]^T$, and let Δx^0 be the direction where it climbs up the tangent plane the fastest.

From space analytic geometry, we know that a pair of unparalleled vector on the tangent plane can uniquely define the plane, and such pair of vector is not difficult to find, as

$$\vec{v}_1 = \left(1, \quad 0, \quad \left.\frac{\partial f(x)}{\partial x_1}\right|_{x=x^0} \right) \tag{6.16}$$

$$\vec{v}_2 = \left(0, \quad 1, \quad \left.\frac{\partial f(x)}{\partial x_2}\right|_{x=x^0} \right) \tag{6.17}$$

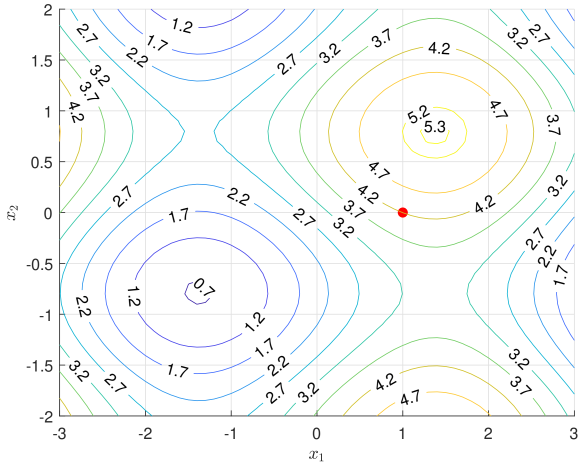

FIGURE 6.4
Contour line of $y = f(x_1, x_2)$.

must be such a pair of vector. This is because (6.16), as shown by the red dashed line in Fig. 6.5, is the tangent of the 2-D intersection of $y = f(x)$ and $x_2 = x_2^0$ shown by the red solid line, therefore must be tangent to the original 3-D surface $y = f(x)$ at x^0. The same applies to (6.17). The tangent plane derived from these two vectors is shown in Fig. 6.6.

The next step is to find a vector $\vec{v}$ on the tangent plane along which y increases the fastest. For convenience, we will do a plane transformation to the tangent plane so that it crosses the origin $(0, 0, 0)$. This can be done by mapping $\left(x_1^k, x_2^k, f(x_1^k, x_2^k)\right)$ to $(0, 0, 0)$. The vector $\vec{v}$ must fulfill the following two conditions: (a) it must be on the tangent plane; (b) it must be perpendicular to the intersection line of the tangent plane and the $y = 0$ plane.

Consider (a). Since the vector is on the tangent plan, it can be represented as a linear combination of (6.16) and (6.17) as

$$\vec{v} \quad = \quad \left(\lambda_1, \quad \lambda_2, \quad \lambda_1 \left.\frac{\partial f(x)}{\partial x_1}\right|_{x=x^0} + \lambda_2 \left.\frac{\partial f(x)}{\partial x_2}\right|_{x=x^0} \right). \tag{6.18}$$

Consider (b). The analytical expression for the tangent plane can be obtained by calculating its normal vector as follows.

$$\begin{aligned} \vec{n} &= \vec{v}_1 \times \vec{v}_2 \\ &= \left(-\left.\frac{\partial f(x)}{\partial x_1}\right|_{x=x^0}, \quad -\left.\frac{\partial f(x)}{\partial x_2}\right|_{x=x^0}, \quad 1 \right) \end{aligned}$$

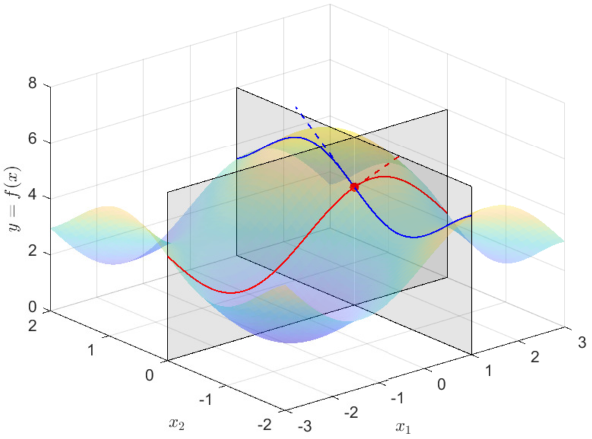

FIGURE 6.5
Plot of vectors given by (6.16) and (6.17).

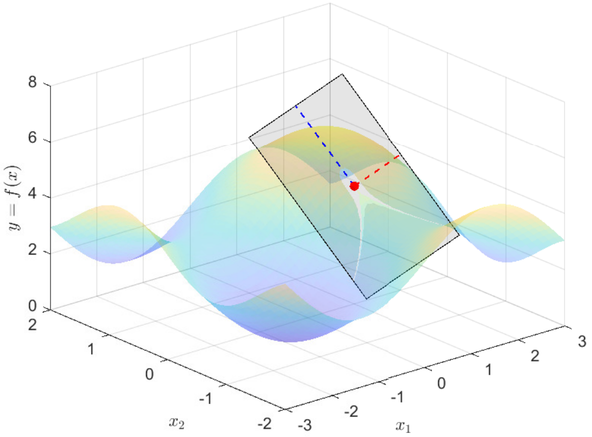

FIGURE 6.6
Formulation of the tangent plane from vectors given by (6.16) and (6.17).

Therefore, the tangent plan is given by (cross origin $(0,0,0)$)

$$y = \left.\frac{\partial f(x)}{\partial x_1}\right|_{x=x^0} x_1 + \left.\frac{\partial f(x)}{\partial x_2}\right|_{x=x^0} x_2$$

And its intersection with plane $y = 0$ is

$$\left.\frac{\partial f(x)}{\partial x_1}\right|_{x=x^0} x_1 + \left.\frac{\partial f(x)}{\partial x_2}\right|_{x=x^0} x_2 = 0 \tag{6.19}$$

Vector $\vec{v}$ must be perpendicular to (6.19). The direction of the intersection can be represented by a vecotr. From (6.19), for example,

$$\vec{v}_{\text{ints}} = \left(\left.\frac{\partial f(x)}{\partial x_2}\right|_{x=x^0}, \ -\left.\frac{\partial f(x)}{\partial x_1}\right|_{x=x^0}, \ 0 \right), \tag{6.20}$$

is a good choice. Vector $\vec{v}_{\text{ints}}$ in (6.20) is perpendicular to $\vec{v}$ in (6.18). From (6.18) and (6.20), equating $\vec{v} \cdot \vec{v}_{\text{ints}} = 0$ gives

$$\lambda_1 \left.\frac{\partial f(x)}{\partial x_2}\right|_{x=x^0} = \lambda_2 \left.\frac{\partial f(x)}{\partial x_1}\right|_{x=x^0} \tag{6.21}$$

Equation (6.21) has infinite number of solutions, resulting in infinite number of $\vec{v}$, all of which point to the same direction. For example, select $\lambda_1 = \left.\frac{\partial f(x)}{\partial x_1}\right|_{x=x^0}$, $\lambda_2 = \left.\frac{\partial f(x)}{\partial x_2}\right|_{x=x^0}$ as the solution. Substituting λ_1 and λ_2 into (6.18) gives

$$\begin{aligned}\vec{v} = \Bigg(& \left.\frac{\partial f(x)}{\partial x_1}\right|_{x=x^0}, \ \left.\frac{\partial f(x)}{\partial x_2}\right|_{x=x^0}, \\ & \left.\frac{\partial f(x)}{\partial x_1}\right|_{x=x^0} \left.\frac{\partial f(x)}{\partial x_1}\right|_{x=x^0} + \left.\frac{\partial f(x)}{\partial x_2}\right|_{x=x^0} \left.\frac{\partial f(x)}{\partial x_2}\right|_{x=x^0} \Bigg).\end{aligned} \tag{6.22}$$

Equation (6.22) gives the guidance of the direction from x^0 to x^1 which can hopefully maximize $f(x^1)$. Therefore, Δx^0 can be obtained as follows.

$$\begin{aligned}\Delta x^0 &= \alpha \begin{bmatrix} \left.\frac{\partial f(x)}{\partial x_1}\right|_{x=x^0} \\ \left.\frac{\partial f(x)}{\partial x_2}\right|_{x=x^0} \end{bmatrix} \\ &= \alpha \left.\nabla f(x)\right|_{x=x^0}\end{aligned} \tag{6.23}$$

where $\alpha > 0$ is an adjustable parameter to determine the progressing rate for each iteration and

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \end{bmatrix} \tag{6.24}$$

is defined as the *gradient* of $f(x)$, which by itself is a vector function of x that has the same dimension with x. In this example, since x is a 2×1 vector, $\nabla f(x)$ is also 2×1. Substituting (6.15) into (6.24) gives

$$\nabla f(x) = \left[\begin{array}{c} 2\cos(x_1) + \frac{1}{2}\cos\left(\frac{x_1}{2} + \pi\right) \\ 2\cos(2x_2) \end{array} \right]. \tag{6.25}$$

Substituting $x^0 = [1, 0]^T$ into (6.25) and (6.23) gives

$$\Delta x^0 = \alpha \left[\begin{array}{c} 0.6418 \\ 2 \end{array} \right]$$

The above procedures can be used to iteratively calculate x^k as follows.

$$\begin{aligned} \Delta x^k &= \alpha \nabla f(x)|_{x=x^k} \\ x^{k+1} &= x^k + \Delta x^k \end{aligned}$$

until $\nabla f(x)|_{x=x^k} \approx 0$ or $f(x^{k+1}) - f(x^k) \approx 0$. Following the same concept, each $f(x^{k+1})$ will be slightly larger than $f(x^k)$ and eventually the maximum value of $f(x)$ and its associated x can be achieved. In this example, calculating x^1 to x^{20} gives the following trajectory of x as shown in Figs. 6.7 and 6.8. The value $\alpha = 0.05$ is used. The maximum $y = 5.32$ can be achieved at $x^{20} = [1.32, 0.77]^T$. Note that this is already a practically good approximation to the actual maximum, as shown in Figs. 6.7 and 6.8 To get a even better approximation, consider using smaller α and increase the iteration time.

6.3.2 Gradient of a Multivariable Function

The gradient of $f(x)$ in (6.24) for the motivating example holds true for general multivariable functions, as long as $f(x)$ is differentiable. The formal definition of gradient is given as follows.

Definition of Gradient:

Consider a scalar-valued differentiable function $y = f(x)$ where $x = [x_1, ..., x_n]^T \in \mathbb{R}^{n \times 1}$. The gradient of $f(x)$ is a vector function of x denoted by $\nabla f(x) \in \mathbb{R}^{n \times 1}$ as follows. The symbol ∇ is called the *nabla symbol*.

$$\nabla f(x) = \left[\begin{array}{c} \frac{\partial}{\partial x_1} f(x) \\ \vdots \\ \frac{\partial}{\partial x_n} f(x) \end{array} \right].$$

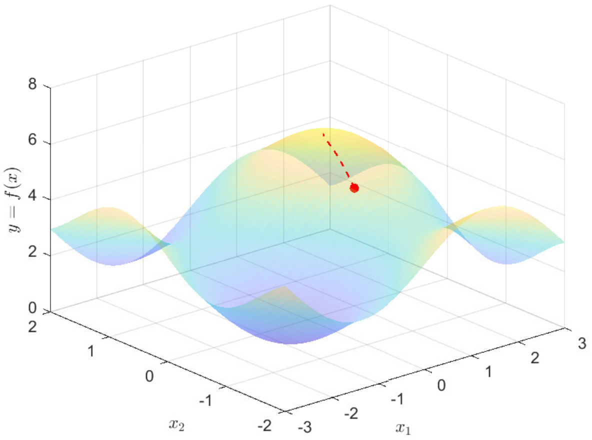

FIGURE 6.7
Trajectory of x until the maximum $y = f(x)$ is achieved.

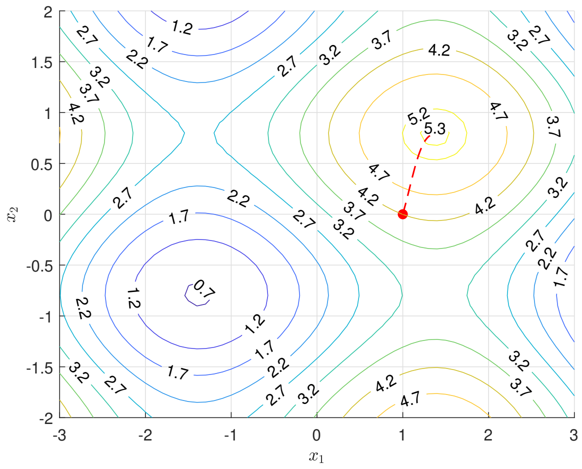

FIGURE 6.8
Trajectory of x until the maximum $y = f(x)$ is achieved on contour line plot.

The gradient of $f(x)$ at point $x = p$ can be calculated by

$$\nabla f(p) = \left. \begin{bmatrix} \frac{\partial}{\partial x_1} f(x) \\ \vdots \\ \frac{\partial}{\partial x_n} f(x) \end{bmatrix} \right|_{x=p}.$$

which can be interpreted as the direction and rate of fastest increase of $f(x)$ at $x = p$.

In case where the direction and rate of fastest decrease of $f(x)$ is required, $-\nabla f(x)$ can be used. Note that $-\nabla f(x)$ is the direction and rate of fastest increase $-f(x)$.

Do note that local maximum/minimum may become an issue while using gradient-based methods to search for maximum/minimum of a function, depending on the function itself and also the initial point x_0 of the iterations.

6.4 Jacobian Matrix

Jacobian matrix is widely used in linear system analysis. For example, it can be used when linearizing a non-linear vector function. The use of Jacobian matrix differs from case to case, thus is not introduced in details here. Only the definition is given as follows.

Definition of Jacobian Matrix:

Consider a vector function $y = f(x)$ where $y = [y_1, ..., y_m]^T \in \mathbb{R}^{m\times 1}$ and $x = [x_1, ..., x_n]^T \in \mathbb{R}^{n\times 1}$. The Jacobian matrix of $f(x)$ is an $m \times n$ matrix, usually denoted by J given by

$$\begin{aligned} J &= \begin{bmatrix} \nabla^T f_1(x) \\ \vdots \\ \nabla^T f_m(x) \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial}{\partial x_1} f_1(x) & \dots & \frac{\partial}{\partial x_n} f_1(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} f_m(x) & \dots & \frac{\partial}{\partial x_n} f_m(x) \end{bmatrix}, \end{aligned}$$

where $f_i(x)$ is the ith element among the m elements in $f(x)$ and $\nabla^T f_i(x)$ is the transpose of $\nabla f_i(x)$.

Sometimes the characteristics of J reveals insights of function $f(x)$, and can be very helpful in evaluating system performance under particular situations.

7

Multiple Integral

CONTENTS

7.1 Motivating Examples 61
7.2 Multiple Integral 66

The integral for scalar input function has been introduced in Chapter 3. The integral of $y = f(x)$ with the lower bound a and upper bound b is denoted by (3.7). It is defined as the limit of sum given by (3.6), and can be interpreted as the area circulated by $x = a$, $x = b$, $y = f(x)$ and $y = 0$ as shown by Fig. 3.4. In practice, the integral can be calculated using (3.8).

In this chapter, the integral for multiple input functions is introduced. Motivating examples are used to illustrate the basic concept and meaning of multiple integral in Section 7.1. The formal definition of multiple integral is given in Section 7.2.

7.1 Motivating Examples

Motivating Example 1

Consider calculating the volume of a cone using integral. The bottom radius and the height of the cone are 1 and 3 respectively.

Figure 7.1 gives a demonstration of the cone in the motivating example. Using similar ideas introduced in Chapter 3, we know that we can think of the cone as a combination of thin cylinders, whose radiuses depend on the vertical position (z-axis position) of the associated cylinder. For example, at $z = 1.5$, the radius of the cylinder is 0.5.

The volume of the cone can be obtained by letting the thickness of each

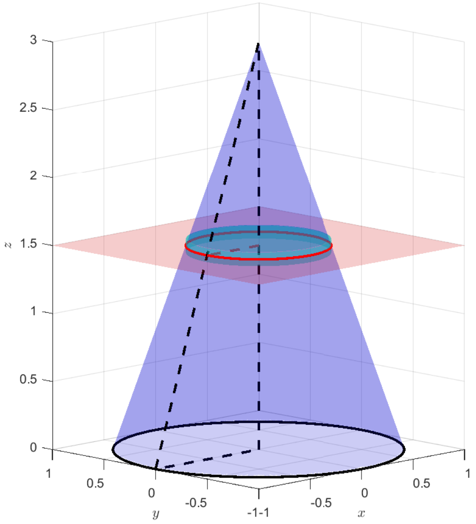

FIGURE 7.1
Calculation of the volume of a cone.

cylinder approaches zero. i.e.,

$$\begin{aligned} V &= \int_0^3 S(z)dz, && (7.1) \\ S(z) &= \pi R(z)^2, && (7.2) \\ R(z) &= \frac{3-z}{3}, 0 \leq z \leq 3, && (7.3) \end{aligned}$$

where $R(z)$ and $S(z)$ are the bottom circle radius and area of the thin cylinder at vertical position z, and $S(z)dz$ can be interpreted as the volume of this thin cylinder.

Substituting (7.2) and (7.3) into (7.1) gives

$$V = \int_0^3 \frac{\pi}{9}(3-z)^2 dz = \frac{\pi}{27}(z-3)^3\Big|_0^3 = \pi,$$

which is consistent with what we learned in primary school: the volume of a cone is one third of its bottom area multiplied by its height.

The above motivating example 1 implies the volume of an object might be formulated as an one-dimensional integration. As a first step, a direction, such as z-axis as given in the motivating example 1, is chosen. Next, imagine using planes to intersect the object. Each plane shall be perpendicular to the selected direction, as given by the red plane in Fig. 7.1. Notice that the red plane is perpendicular to the direction of z-axis. The intersection area shall be a function of the intersection position, as given by (7.2). Finally, the volume can be calculated as the integral of the area function over the direction, as given by (7.1).

In many cases, the calculation of the intersection area can be less simple and intuitive than (7.2). An example is given by the following motivating example.

Motivating Example 2

Consider calculating the volume surrounded by the following surfaces:

$$\begin{aligned} 0 \leq &\quad x \quad \leq 2\pi, \\ 0 \leq &\quad y \quad \leq 2\pi, \\ 0 \leq &\quad z \quad \leq f(x,y) = y\cos(x+y) + 2\pi. \end{aligned} \tag{7.4}$$

We will use the same method adopted from the motivating example 1 to solve motivating example 2. The volume to be calculated is plotted in Fig. 7.2 (only the top surface). The x-axis direction is chosen for the integral in motivating example 2.

It can be spotted soon that motivating example 2 is more complicated than 1 as the intersection area, in this case $S(x)$ as a function of position x, cannot be obtained as intuitively as (7.2).

As an example for illustration, in Fig. 7.2 the red surface is the intersection surface at $x = 2$, and $S(x)$ at $x = 2$ should be the area surrounded by the red line. At a first glance, it seems that $S(x)$ does not have an analytical expression as the red line shape is quite arbitrary.

However, this is not true. The area surrounded by the red line in Fig. 7.2 can be calculated with the knowledge introduced in Chapter 3 as follows. Substituting constant $x = 2$ into (7.4) gives

$$z = f(2,y) = y\cos(y+2) + 2\pi, 0 \leq y \leq 2\pi,$$

which is the analytical expression of the red line (intersecting with the top surface). Therefore, the corresponding area in Fig. 7.2 is

$$\begin{aligned} S(x)|_{x=2} &= \int_0^{2\pi} f(2,y)dy \\ &= \int_0^{2\pi} \left(y\cos(y+2) + 2\pi\right) dy \\ &= \left(y\sin(y+2) + \cos(y+2) + 2\pi y\right)\Big|_0^{2\pi} \\ &= 2\pi\sin(2) + 4\pi^2. \end{aligned}$$

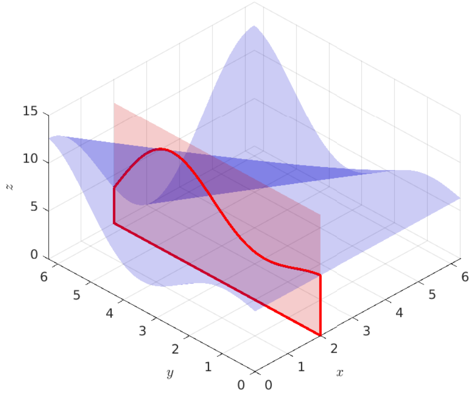

FIGURE 7.2
Calculation of volume of an arbitrary arbitrary project.

Without specifying x as any particular value, $S(x)$ is in general given by

$$\begin{aligned} S(x) &= \int_0^{2\pi} f(x,y)dy && (7.5) \\ &= \int_0^{2\pi} \left(y\cos(x+y)+2\pi\right)dy && (7.6) \\ &= \left(y\sin(x+y)+\cos(x+y)+2\pi y\right)\Big|_0^{2\pi} && (7.7) \\ &= 2\pi\sin(x)+4\pi^2, && (7.8) \end{aligned}$$

where (7.7) is derived from (7.6) by treating x as a constant value. Using (7.5) and (7.8), the volume in motivating example 2 can be calculated as

$$\begin{aligned} V &= \int_0^{2\pi} S(x)dx \\ &= \int_0^{2\pi}\left[\int_0^{2\pi} f(x,y)dy\right]dx && (7.9) \\ &= \int_0^{2\pi}\left[\int_0^{2\pi} \left(y\cos(x+y)+2\pi\right)dy\right]dx \\ &= \int_0^{2\pi} \left(2\pi\sin(x)+4\pi^2\right)dx \\ &= \left(-2\pi\cos(x)+4\pi^2x\right)\Big|_0^{2\pi} \\ &= 8\pi^3. \end{aligned}$$

In (7.9), two integrals are calculated one after another to finally obtain the

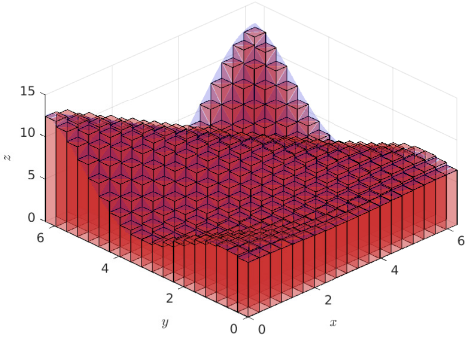

FIGURE 7.3
Calculation of volume of an arbitrary object as sum of cubes.

volume of motivating example 2. There is an alternative way of understanding (7.9) that gives more insights to the problem. From 3, we know that dx and dy are the infinitesimal change along x-axis and y-axis respectively, and the two-step integrals are essentially the calculation of infinite sum of $f(x, y)$ multiplied by Δx and Δy as given in the equation below, with $\Delta x, \Delta y \to 0$. To conclude, (7.9) can be rewritten as

$$\begin{aligned}\int_0^{2\pi}\left[\int_0^{2\pi} f(x,y)dy\right]dx &= \lim_{\Delta x,\Delta y\to 0}\sum_i\left[\sum_j f(x_i,y_j)\Delta y\right]\Delta x \\ &= \lim_{\Delta x,\Delta y\to 0}\sum_{i,j} f(x_i,y_j)\Delta x\Delta y, \qquad (7.10)\end{aligned}$$

where (7.10) represents the sum of a volume of cubes with the bottom area $\Delta x \times \Delta y$ and different height $f(x_i, y_j)$, as demonstrated by Fig. 7.3. With $\Delta x, \Delta y \to 0$, the precise volume of the arbitrary object can be obtained.

Notice that as long as (7.10) exists, the sequence of the summation is irrelevant to the result. For example, the cubes with the same x-axis position can be added together first (as demonstrated by the red cubes in Fig. 7.4). Then the sum of volume of cubes at different x-axis positions are added together. Similarly, it can be done in the other way around. The cubes with the same y-axis position can be added together first (as demonstrated by the green cubes in Fig. 7.4).

Equation (7.10) is denoted by

$$\lim_{\Delta x,\Delta y\to 0}\sum_{i,j} f(x_i,y_j)\Delta x\Delta y = \int_0^{2\pi}\int_0^{2\pi} f(x,y)dxdy, \qquad (7.11)$$

which is an example of double integral. And in this example, we already know

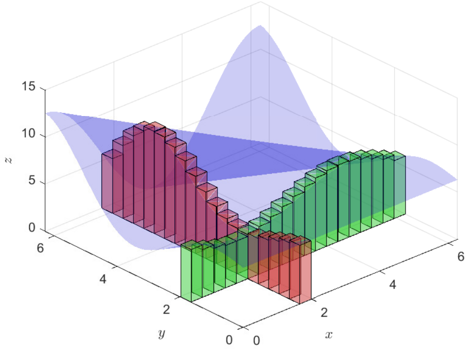

FIGURE 7.4
Summation of cubes in different sequence.

that to solve (7.11), we can use either (7.9) or

$$\int_0^{2\pi}\int_0^{2\pi} f(x,y)dxdy = \int_0^{2\pi}\left[\int_0^{2\pi} f(x,y)dx\right]dy,$$

both shall give the same result as the sequence of integral over x or y is irrelevant as long as (7.11) exists. Notice that for multiple integral sometimes $\iint$, $\iiint$, etc., symbols are used for simplicity when the upper and lower boundary of the integral variables are the same or not given in the equation. Thus, (7.11) can also be written as $\iint_0^{2\pi} f(x,y)dxdy$.

7.2 Multiple Integral

The definition of double integral is given below. Higher order of integral can be achieved by expanding the double integral into a higher dimension hyperspace.

Definition of Double Integral:

Given a function $f(x,y)$, the double integral of f over the rectangle R is defined by

$$\iint_R f(x,y)dA = \lim_{m,n\to\infty}\sum_{i=1}^{m}\sum_{j=1}^{n} f(x_{ij},y_{ij})\Delta A \tag{7.12}$$

if this limit exists. In (7.12), $f(x_{ij}, y_{ij})$ is an arbitrary sample of $f(x, y)$ in the associated infinitesimal area ΔA.

If $f(x, y)$ is continuous on the rectangle $R = \{(x, y) | a \leq x \leq b, c \leq y \leq d\}$, then

$$\begin{aligned} \iint_R f(x,y)dA &= \iint_R f(x,y)dxdy \\ &= \int_a^b \int_c^d f(x,y)dxdy \\ &= \lim_{m,n\to\infty} \sum_{i=1}^{m} \sum_{j=1}^{n} f(x_{ij}, y_{ij})\Delta x \Delta y \end{aligned} \tag{7.13}$$

Equation (7.13) can be interpreted as follows. The rectangular is divided into $m \times n$ infinitesimal "squares", whose length and width given by Δx, Δy respectively. An example is given in Fig. 7.3 where the bottom of each red cube is such a square. The volume of such a cube is then approximated using $f(x_{ij}, y_{ij})\Delta x \Delta y$, where $f(x_{ij}, y_{ij})$ is an arbitrary sample of $f(x, y)$ within the associated square. With the populating of the number of cubes, the right side $f(x_{ij}, y_{ij})\Delta x \Delta y$ may converge (to the volume between $z = f(x, y)$ and $z = 0$ in the given rectangle R). If it indeed converges, then the converged value is denoted by the double integral $\iint_R f(x, y)dxdy$, where $dxdy$ represents the 2-dimensional infinitesimal square $\Delta x \Delta y$ when they both approach zero.

Equation (7.12) can be expanded to hyperspace for multiple integral with more than 2 variables. In practice, there is no limit to the maximum order of integral in an equation.

8

Applications

CONTENTS

8.1 Neural Network Back-propagation 69
    8.1.1 Perceptron 69
    8.1.2 A Multi-layer Perceptrons Model 69
    8.1.3 Training and Testing 70
8.2 Bayesian Inference 70

Two examples are given in this section for partial differential and multiple integral applications respectively.

Section 8.1 introduces back-propagation for a conventional multi-layer perceptrons artificial neural network. Back-propagation is a key procedure where the neural network "learns" from labeled training samples for pattern recognition.

Section 8.2 introduces Bayesian inference, a widely used method for updating the probability of a hypothesis using Bayes theorem.

8.1 Neural Network Back-propagation

The back-propagation of a conventional multi-layer artificial neural network (ANN) is used as an example to illustrate the use of partial differential.

For the convenience of the reader, preliminary knowledge of the ANN, such as the concept of a perceptron, is introduced in 8.1.1. The multi-layer perceptrons model used in the example is introduced in 8.1.2. Finally the ANN is trained and tested using the training and testing sets, and the results are given in 8.1.3.

8.1.1 Perceptron

“nobreak

8.1.2 A Multi-layer Perceptrons Model

“nobreak

8.1.3 Training and Testing

“nobreak

8.2 Bayesian Inference

Part III

Differential Equation

9

Ordinary Differential Equation

CONTENTS

9.1 Definition 73
9.2 Canonical Forms and Solutions 73

“nobreak

9.1 Definition

“nobreak

9.2 Canonical Forms and Solutions

10

Partial Differential Equation

CONTENTS

10.1 Definition 75
10.2 Canonical Forms and Solutions 75

“nobreak

10.1 Definition

“nobreak

10.2 Canonical Forms and Solutions

11

Applications

CONTENTS

11.1 Circuit Transient Analysis 77
11.2 Finite Element Analysis 77

“nobreak

11.1 Circuit Transient Analysis

“nobreak

11.2 Finite Element Analysis

Part IV

Calculus of Variations

12

Functional

CONTENTS

12.1 Motivating Example 81
12.2 Functional 82
    12.2.1 Definition 82
    12.2.2 Finite Differences Approximation 83
    12.2.3 Function Space 84
    12.2.4 Normed Linear Space and Norm of Function 84
    12.2.5 Functional Continuity and Linearity 85
12.3 Variation of Functional 86

Functional refers to a mapping of functions or curves that belong to a certain set to a definite number. Details are introduced in this chapter.

12.1 Motivating Example

One of the most famous problems solved by functional and calculus of variation is the brachistochrone problem, which is described in the box below.

Apparently, the brachistochrone problem is an optimization problem where the objective is to find a curve that minimizes a cost function, in this case, the travelling time of the particle. This differs from a conventional optimization problem where the target is to find the optimum values for a limited number of variables.

The curve, or equivalently the analytical function $y = f(x)$ that describes the curve, contains infinite number of “pixels”. From that sense, the problem can be taken as an optimization problem with infinite number of degrees. Therefore, the problem cannot be converted into a conventional optimization problem easily.

Brachistochrone Problem

Let A and B be two fixed points in the space that are not aligned either vertically or horizontally. Without loss of generality, let A be higher than B. Draw a curve that joins A and B. A particle is released from A and it slides along the curve traveling from A to B driven by gravity.

Fraction effect is ignored, and hence the velocity of the particle at any given position on the curve is purely decided by the vertical position of the particle relevant to A (speed) and the tangent of the curve at the position (direction).

A demonstration plot is given by Fig. 12.1.

The brachistochrone problem tries to find the curve that minimizes the time it takes for the particle to travel from A to B. The objective is to find the analytical solution $y = f(x)$ that describes such curve.

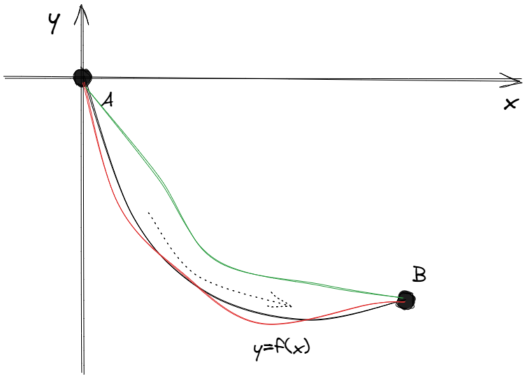

FIGURE 12.1
Brachistochrone problem description.

12.2 Functional

Definition of functional as well as its relationship with the associated finite differences approximation is explained as follows.

12.2.1 Definition

Functional is a map of a function to a number. It can be regarded as an extension to function, where the input is not a variable or a vector, but a function. Functional has many practical applications in data analysis, mechanics, geometry, control systems, etc. Many optimization problems are also defined using functional.

Let $y(x)$ be a function of x. Let $J[y]$ be a functional of x, y and values derived from them. For example, $J[y]$ may look like

$$J = \int_a^b \sqrt{1+y\prime^2}dx$$

which is the length of curve $y(x)$ from $x = a$ to $x = b$. The study of the properties of $J[y]$ with different choice of $y(x)$ is known as the **calculus of functional**.

Calculus of functional is often difficult to solve. **Calculus of variation** is a special case of calculus of functional that only studies the choice of y that maximizes or minimizes J. Even with that been said, calculus of variation is still difficult to solve in general, and the analytical solution of $y(x)$ exists for only certain forms of J. This notebook only considers a small subset of the problem where J follows the following form

$$\begin{aligned} y &= y(x) \\ J &= \int_a^b F\left(x, y, y\prime\right)dx \end{aligned} \tag{12.1}$$

Notice that as far as this notebook concerns, the functional (12.1) must possess **localization property**, where if the curve $y(x)$ is split into segments whose functional calculated separately, the sum of the values of the functional would be equal to the functional of the whole curve. The form in (12.1) satisfies this property. This does not generally apply if J is defined arbitrarily. For example, consider

$$J = \frac{\int_a^b x\sqrt{1+y\prime^2}dx}{\int_a^b \sqrt{1+y\prime^2}dx}$$

which does not follow (12.1) and it is a non-local functional.

12.2.2 Finite Differences Approximation

An intuitive way of solving the calculus of variation is as follows. Let x_i, $i = 0, ..., n+1$ be samples uniformly taken from $[a, b]$ where $x_0 = a$ and $x_{n+1} = b$. Let $y_i = y(x_i)$. With localization property, (12.1) can be approximated by the following finite differences

$$J[y] \approx \sum_{i=1}^{n+1} F\left(x_i, y_i, \frac{y_i - y_{i-1}}{h}\right) \tag{12.2}$$

where $h = x_{i+1} - x_i$. When $n \to \infty$, (12.2) converges to (12.1). The above converts the calculus of variation problem into a finite differences problem.

Instead of looking for a continuous curve $y(x)$, the problem is now looking for a vector $[y_1, \dots, y_n]$ that maximizes or minimizes J. Each $y(x)$ in the domain of the functional corresponds with a vector $[y_1, y_2, \dots, y_n]$ which can be taken as a coordinate in the n-dimensional space.

12.2.3 Function Space

A function $y(x)$ has a domain that defines the set of possible input variables of x. Similarly, a functional $J[y]$ has a **function space** that defines the set of possible choice of curves of y.

The function space is determined by both the physical law behind the problem as well as the formulation of the functional. For example, consider functional

$$J[y] \quad = \quad \int_a^b F\left(x, y, y'\right) dx$$

It is rather natural to assume the function space to be any function defined on the closed interval $[a, b]$ that is continuous and has (continuous) 1-st order derivative.

Notice that with finite differences approximation (12.2), each y in the function space corresponds with a vector $[y_1, \dots, y_n]$, or a “dot” in the n-dimensional space. This may help with understanding the concept of norm which will me introduced in the remaining of this section.

12.2.4 Normed Linear Space and Norm of Function

It is critical to define the “closeness”, i.e., the distance of two functions in a function space. For example, in Fig. 12.1, it is natural to think that the red curve is closer to the black curve than the green curve.

The distance of two points in Euclidean space can be easily defined. Using that analogy, the most convenient way to define the distance of two functions in the function space is to use the norm. To better understand norm, the concept of a normed linear space is introduced as follows.

Consider a set (space) A, and elements in that space $x, y, z, ... \in A$. Space A is a **linear space** if addition and multiplication operations are defined, and there are spacial elements $0 \in A$ and $1 \in A$ in the set, and the following criteria are met. More details are given in abstract algebra relevant sections in the another notebook.

$$\begin{aligned}
x+y &= y+x \\
(x+y)+z &= x+(y+z) \\
x+0 &= x \\
\forall x \Rightarrow \exists(-x), (-x)+x &= 0 \\
1 \cdot x &= x \\
\alpha(\beta x) &= (\alpha\beta)x \\
(\alpha+\beta)x &= \alpha x + \beta x \\
\alpha(x+y) &= \alpha x + \alpha y
\end{aligned}$$

where $\alpha, \beta \in \mathbb{R}$.

Further more, A is **normed** if for $\forall x \in A$, there is a non-negative number denoted by $\|x\|$ so that

$$\begin{aligned}
\|x\| = 0 &\Leftrightarrow x = 0 \\
\|\alpha x\| &= |\alpha|\|x\| \\
\|x+y\| &\leq \|x\| + \|y\|
\end{aligned}$$

and A becomes a **normed linear space**.

In a normed linear space, the distance between two elements x, y is defined as $\|x-y\|$. Using norm, the following distances are defined based on the property of the function space.

Consider function space $\mathcal{C}(a,b)$ that consists of all continuous $y(x)$ defined on a closed interval $[a,b]$. Let y^1, y^2 be two functions in the function space. The following norm is often used to define the difference of y^1 and y^2.

$$\left\|y^1(x) - y^2(x)\right\|_0 = \max_{a \leq x \leq b} |y^1(x) - y^2(x)| \tag{12.3}$$

Consider function space $\mathcal{D}_1(a,b)$ that consists of all continuous and 1-st order differentiable functions $y(x)$ defined on a closed interval $[a,b]$. Let y^1, y^2 be two functions in the function space. The following norm is often used to define the difference of y^1 and y^2.

$$\begin{aligned}
\left\|y^1(x) - y^2(x)\right\|_1 &= \max_{a \leq x \leq b} |y^1(x) - y^2(x)| \\
&\quad + \max_{a \leq x \leq b} |y^{1\prime}(x) - y^{2\prime}(x)|
\end{aligned} \tag{12.4}$$

Consider function space $\mathcal{D}_n(a,b)$ that consists of all continuous and up to n-th order differtentiable functions $y(x)$ defined on a closed interval $[a,b]$. Let y^1, y^2 be two functions in the function space. The following norm is often used to define the difference of y^1 and y^2.

$$\left\|y^1(x) - y^2(x)\right\|_n = \sum_{i=0}^{n} \max_{a \leq x \leq b} |y^{1(i)}(x) - y^{2(i)}(x)| \tag{12.5}$$

12.2.5 Functional Continuity and Linearity

A function $y(x)$ is said to be continuous at x_0 if $y(x) - y(x_0)$ can be as small as desired provided that $x - x_0$ is small. The similar idea applies to functional continuity.

A functional $J[y]$ is said to be a **continuous functional** at y^0 if for any small ϵ, there is always a corresponding δ, so that

$$J[y] - J[y_0] \quad < \quad \epsilon$$

provided that $\left\|y - y^0\right\| < \delta$, where $\|\cdot\|$ is one of the norms introduced in (12.3), (12.4) or (12.5).

Functional continuity guarantees that "similar" functions y^1 and y^2 in the function space result in close $J[y_1]$ and $J[y_2]$, where "similar" is defined by the norm of the function that reflects the maximum differences of two functions in magnitude or in differentials magnitudes.

Functional continuity on functional space $\mathcal{C}$ or $\mathcal{D}_1$ is often the minimum prerequisite where calculus of variation (to be introduced in the remaining sections) can be applied.

Continuous functional $J[y]$ is said to be a **continuous linear functional** if the following criteria are met.

$$\begin{aligned} J[\alpha y] &= \alpha J[y] \\ J[y_1 + y_2] &= J[y_1] + J[y_2] \end{aligned}$$

where α is a scalar.

Notice that the linearity of $J[y]$ depends not on how y is constructed as a function of x, but on how $J[y]$ is constructed as a function of y. It can be proved that a functional of the following form

$$J[y] \quad = \quad \int_a^b \left(\alpha_0(x)y(x) + \alpha_1(x)y'(x) + \ldots + \alpha_n y^{(n)}(x)\right) dx \qquad (12.6)$$

where $\alpha_i(x)$ is a fixed function, defines a linear functional on $\mathcal{D}_n$.

12.3 Variation of Functional

Let $J[y]$ be a functional defined on some normed linear space. Let

$$\Delta J[h] \quad = \quad J[y + h] - J[y] \qquad (12.7)$$

be the increment corresponding to h of y. Notice that $J[y]$ is not necessarily linear. If $J[y]$ is indeed linear, (12.7) gives

$$\Delta J[h] \quad = \quad \varphi(h)$$

depends on only h but not y. Here, $\varphi(h) = J[h]$ is a linear functional of h. If $J[y]$ is nonlinear, (12.7) can be formulated as

$$\Delta J[h] \quad = \quad \phi(h) + \varepsilon \, \|h\|$$

where $\varphi(h)$ is a linear functional of h, playing the principal linear part of $J[y]$, and $\varepsilon \, \|h\|$ the remainder, and $\varepsilon \to 0$ as $\|h\| \to 0$. In this case, $J[y]$ is said to be differentiable. The principal linear part of the increment $\phi(h)$ is "almost" $\Delta J[h]$ but differs from it by an infinitesimal of order higher than 1 relative to $\|h\|$, and it is known as the **variation of functional** of $J[y]$, often denoted by $\delta J[h]$.

Part V

Numerical Analysis

13

Brief Introduction to Numerical Analysis

CONTENTS

13.1 Motivation 91
13.2 Challenges 91

Numerical analysis extends beyond the scope of calculus, yet it remains pertinent to include it within a calculus-focused notebook. This is because many numerical analysis algorithms derive their principles from calculus. Textbooks on numerical analysis, such as [1] (currently in its 10th edition), often begin with a "review of calculus" as its first chapter, underscoring its foundational importance.

13.1 Motivation

Analytical equations alone sometimes fail to adequately describe a variable or a model. Consider cases where we must solve $f(x) = 0$ for x, but no analytical expression for x exists due to the complexity of $f(x)$. Similarly, describing the behavior of a model through $f(x)$ accurately using analytical expression may be impractical if the true model is too complex to calibrate accurately. In these instances, analytical solutions reach their limits.

Numerical solutions provide a viable alternative to analytical methods. Although they are typically more computationally intensive and may offer less insight than their analytical counterparts, numerical solutions are indispensable in numerous engineering problems where analytical methods fall short.

Numerical analysis focuses on the effectiveness and efficiency of these numerical solutions across a diverse array of problems.

13.2 Challenges

In contrast to analytical solutions which are often presented through equations, numerical solutions usually involve iterative or recurrent algorithms. These algorithms approximate the analytical solution, aiming to balance accuracy with computational feasibility. Numerical analysis addresses several critical aspects of these solutions:

Convergency and Robustness of the Algorithm

Iterative algorithms are generally effective but may fail to converge under certain conditions. Understanding and defining the convergence criteria is essential for ensuring the reliability and robustness of numerical algorithms.

Approximation Error

Numerical solutions commonly employ approximations, leading to potential discrepancies between the numerical and true solutions. Numerical analysis aims to quantitatively assess and minimize these errors.

Computational Burden and Speed of Convergence

Numerical computations can be time-consuming, and complex algorithms may require substantial computational resources, sometimes exceeding practical limits. This computational demand can significantly constrain the application of numerical solutions, particularly in real-time scenarios.

14

One-Variable Equation

CONTENTS

14.1 General Problem Formulation 93
14.2 Solution 94
  14.2.1 Bisection Method 94
  14.2.2 Fixed Point Iteration Method 95
  14.2.3 Newton's method 97
  14.2.4 Secant Method 98
  14.2.5 Muller's Method 99
14.3 Convergence Speed and Error 99
  14.3.1 Order of Convergence 99
  14.3.2 Multiple Root 100
  14.3.3 Convergence Acceleration 101
14.4 Root for Polynomial 101

There are variety of ways to solve the scalar function $f(x) = 0$ for scalar independent variable x, the most intuitive of which being deriving its analytical solution in the explicit form. However, in many cases the explicit analytical solution of x may not be achievable due to the complexity of $f(x)$. To address this problem, many numerical analysis based methods have been developed to obtain (an approximation of) the solution effectively and efficiently.

This chapter discusses these numerical methods as well as their convergence and computational burden.

It is worth mentioning that many methods introduced in this chapter, such as Newton's method, can be easily expanded to solve vector functions with vector independent variables. Nevertheless, for the convenience of the illustration, only scalar function with scalar input is considered in this chapter.

14.1 General Problem Formulation

Let $f(x) = 0$ be a scalar function, where $f(x)$ is continuous. There is at least one solution to $f(x) = 0$, namely p, and $p \in [a, b]$. The target is to find p. A demonstrative plot is given in Fig. 14.1. In the example, function $f(x) = 0$ has two solutions within $[0, 10]$. The target is to find any one of them.

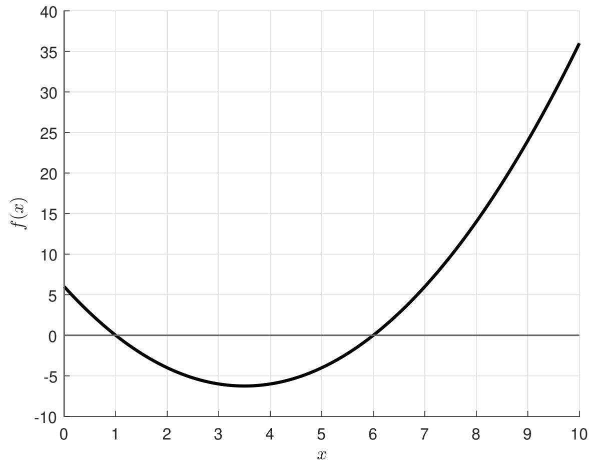

FIGURE 14.1
A demonstrative example of solving $f(x) = 0$ using numerical methods.

Notice that there might be other solutions to $f(x) = 0$ within $[a, b]$ than p. Useful it may be to find all solutions, however, in the scope of our discussion we look for only one solution.

14.2 Solution

A scalar function can be solved numerically via the following methods, each with some pros and cons.

14.2.1 Bisection Method

Bisection method, also known as the binary search method, is inspired by the *intermediate value theorem.* So long as we can find an interval $[a, b]$ so that $f(a)f(b) < 0$, we know that there must be such a solution $p \in [a, b]$ that $f(p) = 0$. If we can narrow down the interval, we can approach the solution.

Intermediate Value Theorem

Let $f(x)$ be a continuous function whose domain contains the interval $[a, b]$. Without loosing generality, let us assume that $f(a) < f(b)$. Then $\forall c \in [f(a), f(b)]$, $\exists x_0 \in [a, b]$, so that $f(x_0) = c$.

The algorithm is summarized as follows. Let $f(x)$ be a continuous function on $[a, b]$, and $f(a)f(b) < 0$. Do the following to find a solution $f(p) = 0$ within $[a, b]$.

1. Let $a_1 = a$, $b_1 = b$.
2. Let $p_i = \frac{a_i+b_i}{2}$.
3. Calculate $f(p_i)$. If $f(p_i) = 0$, then p_i is a solution. Otherwise:
    - If $f(p_i)f(a_i) < 0$, let $a_{i+1} = a_i, b_{i+1} = p_i$.
    - If $f(p_i)f(b_i) < 0$, let $a_{i+1} = p_i, b_{i+1} = b_i$.
4. Check stop criteria as follows:
    - If the solution p_i is found, return p_i;
    - If the interval $b_{i+1} - a_{i+1}$ is less than a threshold, return either a_{i+1} or b_{i+1};
    - If the number of iteration exceeds the maximum iteration limit, return either a_{i+1} or b_{i+1} together with a warning flag;
    - Otherwise, iterate from Step 2.

Bisection method is intuitive and it can definitely find the solution. The down side of the method is that it takes many iterations to converge, hence not very efficient.

14.2.2 Fixed Point Iteration Method

A *fixed point* of a function $f(x)$ is such $x = x_0$ that $f(x_0) = x_0$. The problem of solving an equation can be converted into finding a fixed point of a function. Consider solving $f(x) = 0$ for its solution p. Let

$$g(x) = x + af(x) \tag{14.1}$$

where $a \neq 0$ can be any value we choose, for example $a = 1$. Clearly, a fixed point p of $g(x)$ must be a solution to $f(x)$ because $g(p) = p$ according to the definition and substituting it into (14.1) gives

$$p = p + af(p) \tag{14.2}$$

hence $f(p) = 0$ for $a \neq 0$.

The remaining of this section discusses when the fixed point exists and how to find one.

Fixed points of a function can be easily identified in the plot. A fixed point of $g(x)$ is its intersection with $y = x$ as shown in Fig. 14.2. Function $g(x) = x^2 - 2$ is given by the solid line and $y = x$ the dashed line. Their two intersections, $p_1 = -1$ and $p_2 = 2$, are the fixed point of $g(x)$.

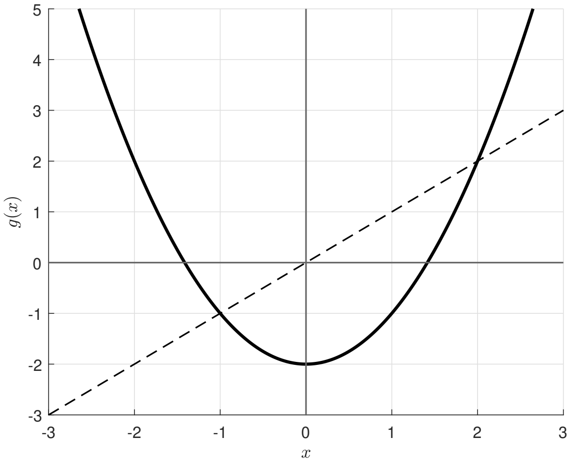

FIGURE 14.2
Function $g(x) = x^2 - 2$ and its two fixed points, $p_1 = -1$ and $p_2 = 2$.

The following criterion can be used to find an interval for a fixed point:

- For continuous function $g(x)$, if for an interval $[a, b]$, $\forall x \in [a, b]$, $g(x) \in [a, b]$, then there is at least one fixed point of $g(x)$ in interval $[a, b]$.
- Furthermore, if the derivative $g'(x)$ exists in (a, b) and $\forall x \in (a, b)$, $|g'(x)| < 1$, then the fixed point is unique.

The proof is fairly straight forward and are not given here.

The fixed point of a function might be found via fixed point iteration. The idea is to use a series $\{p_n\}$ to approximate the fixed point p with $n \to \infty$. The steps are summarized below.

1. Select $p_1 \in [a, b]$
2. Calculate $p_{i+1} = g(p_i)$
3. Check stop criteria as follows:
    - If $p_{i+1} - p_i$ is less than a threshold, return p_{i+1};

- If the number of iteration exceeds the maximum iteration limit, return p_{i+1} together with a warning flag;
- Otherwise, iterate from step 2.

This method is illustrated in Fig. 14.3. It can be seen from the figure how series $\{p_n\}$ approaches the fixed point p.

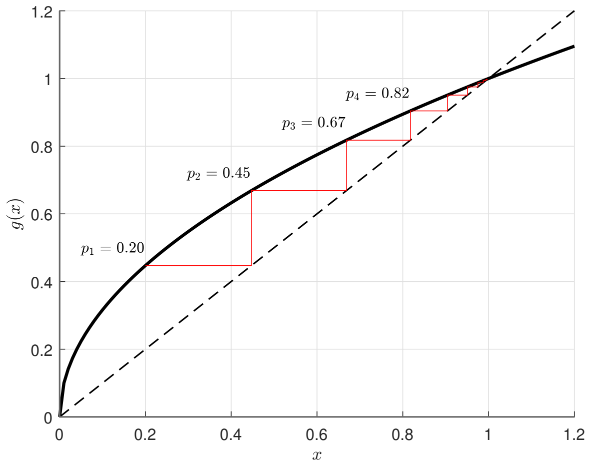

FIGURE 14.3
Approaching fixed point using fixed point iteration.

Notice that unlike bisection method, the aforementioned fixed point iteration method does not necessarily converge, and the key is the gradient $|g'(x)|$ within the interval (a, b). To be more precise,

- If for $g(x)$ defined in interval $[a, b]$, if $g(x) \in [a, b]$ and $|g'(x)| < 1$ in (a, b), then the aforementioned fixed point iteration method converges.

and this can be interpreted easily from a plot like Fig. 14.3.

14.2.3 Newton's method

Newton's method, or Newton-Raphson method, is one of the most popular methods to solve equations numerically. It is intuitive and in most applications it converges fairly easily and fast. Broadly speaking, Newton's method is a special case of the fixed point iteration as it also construct a series $\{p_n\}$ that approaches to the solution of $f(x) = 0$. Unlike (14.1), the series looks like

$$g(x) = x - \frac{f(x)}{f'(x)} \tag{14.3}$$

given that $f'(x) \neq 0$ in the concerned interval (a, b). A fixed point of (14.3), i.e. $g(p) = p$, obviously satisfies $f(p) = 0$ and it can be proved by substituting $g(p) = p$ into (14.3).

From (14.3), consider constructing the series as follows

$$p_{i+1} = p_i - \frac{f(p_i)}{f'(p_i)} \tag{14.4}$$

Newton's method iterates (14.4) to approach the solution. A demonstrative Fig. 14.4 is given below. It can be seen from Fig. 14.4 that in this example the iteration in (14.4) quickly converges the solution by tracing the gradient of $f(x)$.

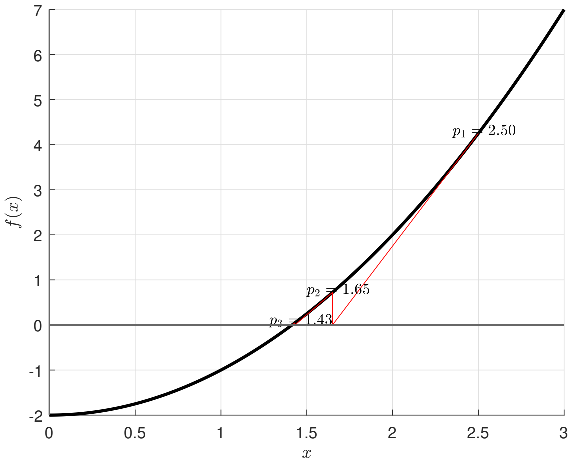

FIGURE 14.4
A demonstration of using Newton's method to solve an equation.

Indeed, with the same number of iterations, Newton's method usually gets to the solution faster than the bisection and the conventional fixed point iteration methods. The more "linear" $f(x)$ is around the solution, the faster and more accurately Newton's method would converge. As long as $f'(x) \neq 0$ at the solution, there is always a neighbourhood around the solution where Newton's method converges.

The only drawback of Newton's method, comparing with the other methods introduced so far, is that it requires the calculation of $f'(x)$ in the iteration. This may not be a big issue if $f'(x)$ can be calculated easily (for example, when its analytical form is available), but it may pose a problem if this is not the case.

14.2.4 Secant Method

The secant method might be used as an alternative to the Newton's method. Instead of (14.4), the secant method uses the following series

$$p_{i+1} = p_i - \frac{f(p_i)(p_i - p_{i-1})}{f(p_i) - f(p_{i-1})}$$

with two closed initial guesses p_1 and p_2 in the interval. The idea behind secant method is simple: it uses the secant of two consecutive samples, p_i and p_{i-1}, to approximate the gradient $f'(p_i)$.

The secant method is a little bit slower than the Newton's method in terms of convergence. Yet, it can be useful when $f(x)$ is significantly easier to calculate than $f'(x)$.

14.2.5 Muller's Method

Muller's method is an extension to the secant method. Recall in the secant method, two initial points p_1 and p_2 are selected as the initial samples. In the essence, it fits a first-order polynomial $P(x) = ax + b$ that passes through $p_1, f(p_1)$ and $(p_2, f(p_2))$. The value of a, b can be calibrated and the solution to $P(x) = 0$ can be calculated analytically, which gives p_3. If $f(p_3) = 0$, it returns the solution. Otherwise, it iterate the process with $(p_2, f(p_2))$ and $(p_3, f(p_3))$ now becoming the new samples to form the first-order polynomial.

Instead of using first-order polynomial $P(x) = ax + b$ to approximate the original function $f(x)$ near its solution, Muller's method uses second-order polynomial $P(x) = ax^2+bx+c$ for the same trick. To calibrate the parameters, three samples, p_1, p_2 and p_3 are required instead of two samples in the case of the secant method.

14.3 Convergence Speed and Error

The convergency of the methods introduced earlier have been briefly discussed in the previous section. This section digs deeper into the convergency criteria and speed of the methods, and performs error analysis to the methods. New algorithms will be proposed on top of the methods to improve them.

14.3.1 Order of Convergence

Let $\{p_n\}$ be a series that converges to p. To measure the speed of convergence of the series, define the order of convergence α and the asymptotic error constant λ as follows.

$$\lim_{i\to\infty} \frac{|p_{i+1} - p|}{|p_i - p|^{\alpha}} = \lambda \tag{14.5}$$

In (14.5), the larger the value of α, the faster the convergence. When $\alpha = 1$ and $\lambda < 1$, the series is called linear convergence. When $\alpha = 2$ and $\alpha = 3$, the series is called quadratic convergence and cubic convergence, respectively. With everything otherwise the same, we usually prefer a series with a higher order of convergence when using it in the fixed point iteration.

Let g be the recurrence relation of series $\{p_n\}$, i.e., $p_{i+1} = g(p_i)$. The series converges to p. The following statements are true. The proof is not given in this notebook.

- If $g'(p) \neq 0$, then $\{p_n\}$ is linear convergence with asymptotic error constant $|g'(p)|$.
- If $g'(p) = 0$, then there is a neighbourhood of p where $\{p_n\}$ is at least quadratic convergence. Furthermore, let the upper bound $|g''(p)| < M$, then

$$\lim_{i\to\infty} \frac{|p_{i+1} - p|}{|p_i - p|^2} < \frac{M}{2}$$

From (14.1) and (14.3), we know that while the conventional fixed point iteration is linear convergence, the Newton's method is at least quadratic convergence. This explains why Newton's method converges faster than the conventional fixed point iteration.

14.3.2 Multiple Root

When the original equation $f(x) = 0$ has multiple roots at $x = p$, i.e.,

$$f(x) = (x - p)^m q(x)$$

with $\lim_{x\to p} \neq 0$ and $m \geq 2$, the methods introduced so far may not work efficiently.

The multiple roots at $x = p$ can be observed by checking the gradient $f'(x)$ at p. We can prove that

- For a continuous function $f(x)$ defined in $[a, b]$, $p \in (a, b)$ is its single root if and only if $f(p) = 0$ and $f'(p) \neq 0$.
- For a continuous function $f(x)$ defined in $[a, b]$, $p \in (a, b)$ is its m multiple roots if and only if $0 = f(p) = f'(p) = ... = f^{(m-1)}(p)$ and $f^{(m)}(p) \neq 0$.

For Newton's method, on may notice from Fig. 14.4 that if there are multiple roots and $f'(p) = 0$ at the solution, though Newton's method still converges, the convergence may slow down. The intuition is indeed correct. Newton's method is usually quadratic convergence when $f'(p) \neq 0$. However, when $f'(p) = 0$, it may not be the case.

If an equation $f(x) = (x-p)^m q(x)$ has m multiple roots at p, one way to handle it is to define

$$\mu(x) = \frac{f(x)}{f'(x)}$$

where it can be proved that $\mu(x)$ has a single root at $x = p$. Then, apply Newton's method on $\mu(x)$ to get the solution p.

14.3.3 Convergence Acceleration

Aitken Δ^2 method is used to form a new series $\{\hat{p}_n\}$ based on a linear convergent series $\{p_n\}$, where both series converges to p but $\{\hat{p}_n\}$ converges faster.

Aitken Δ^2 method defines

$$\hat{p}_i = p_i - \frac{(p_{i+1} - p_i)^2}{p_{i+2} - 2p_{i+1} + p_i}$$

The idea is to calculate p_i normally, and while doing that also calculate $\hat{p}_i$ using p_i, p_{i+1} and p_{i+2}.

Aitken Δ^2 method can be used to accelerate fixed point iteration. The idea is as follows.

1. Calculate p_1, $p_2 = g(p_1)$ and $p_3 = g(p_2)$ normally.
2. Calculate $\hat{p}_1$ using p_1, p_2 and p_3.
3. Overwrite p_1 with $\hat{p}_1$, and repeat from Step 1.

and it is known as the Steffensen method.

14.4 Root for Polynomial

Polynomial is a special type of a function that looks like

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0$$

where $a_n \in \mathbb{R}$ and $a_n \neq 0$. Variable n is known as the order of the polynomial. In this case, $f(x)$ is often denoted as $P(x)$. Notice that $P(x) = 0$ by concept is also a polynomial with no order, but it is often not an interest of study. For the remaining, we will only look at polynomials with $n \geq 1$.

It can be proved (requiring theory from complex analysis) that:

- Polynomial equation $P(x) = 0$ with $n \geq 1$ has at least one root. The root can be real number or complex number.

- Polynomial equation $P(x) = n$ with the order of n can be uniquely decomposed into the following form:

  $$P(x) = a_n(x - x_1)^{m_1}(x - x_2)^{m_2} \ldots (x - x_k)^{m_k}$$

  where $\sum m_i = n$, and x_i being its m_i multiple root.

- For two polynomials both less or equal to nth order, $P(x)$ and $Q(x)$, if there are distinct values $x_1, x_2, ..., x_k$ where $P(x_i) = Q(x_i)$, $i = 1, ..., k$, and $k > n$, then $P(x) = Q(x)\ \forall x$.

Finding the numerical solution for a general equation $f(x) = 0$ using different methods, such as the Newton's method, has been introduced in earlier sections. When comes to polynomial equation $P(x) = 0$, we can further improve the algorithms for better computation efficiency.

Given a polynomial $P(x)$ and a specified variable x_0, $P(x)$ can be decomposed as follows

$$P(x) = (x - x_0)Q(x) + b_0 \tag{14.6}$$

where

$$\begin{aligned} Q(x) &= b_n x^{n-1} + b_{n-1}x^{n-2} + \ldots + b_2 x + b_1 \\ b_n &= a_n \\ b_k &= a_k + b_{k+1}x_0, k = n-1, n-2, \ldots, 1, 0 \end{aligned}$$

The proof is brute-force and are neglected here.

Obviously, from (14.6) $b_0 = P(x_0)$. The decomposition provides an efficient way to calculate $P(x_0)$, which is usually faster than substituting x_0 into $P(x)$ directly and calculate $a_i x^i$ for all the terms. Furthermore, we know from (14.6) that $P'(x_0) = Q(x_0)$. This helps with speeding up the Newton's method to solve $P(x) = 0$. This is known as the Horner method.

15

Interpolation

CONTENTS

15.1 Lagrange Interpolation 103

Interpolation studies the way to construct an analytical function $y = f(x)$ from a series of samples (x_i, y_i).

15.1 Lagrange Interpolation

Bibliography

[1] Richard L Burden and J Douglas Faires. *Numerical analysis.* Brooks Cole, 1997.

[2] James Stewart. *Calculus Metric Version Eighth Edition.* Cengage Learning, 2015.

[3] Gibert Strang. *Calculus.* Wellesley-Cambridge Press, 1991.