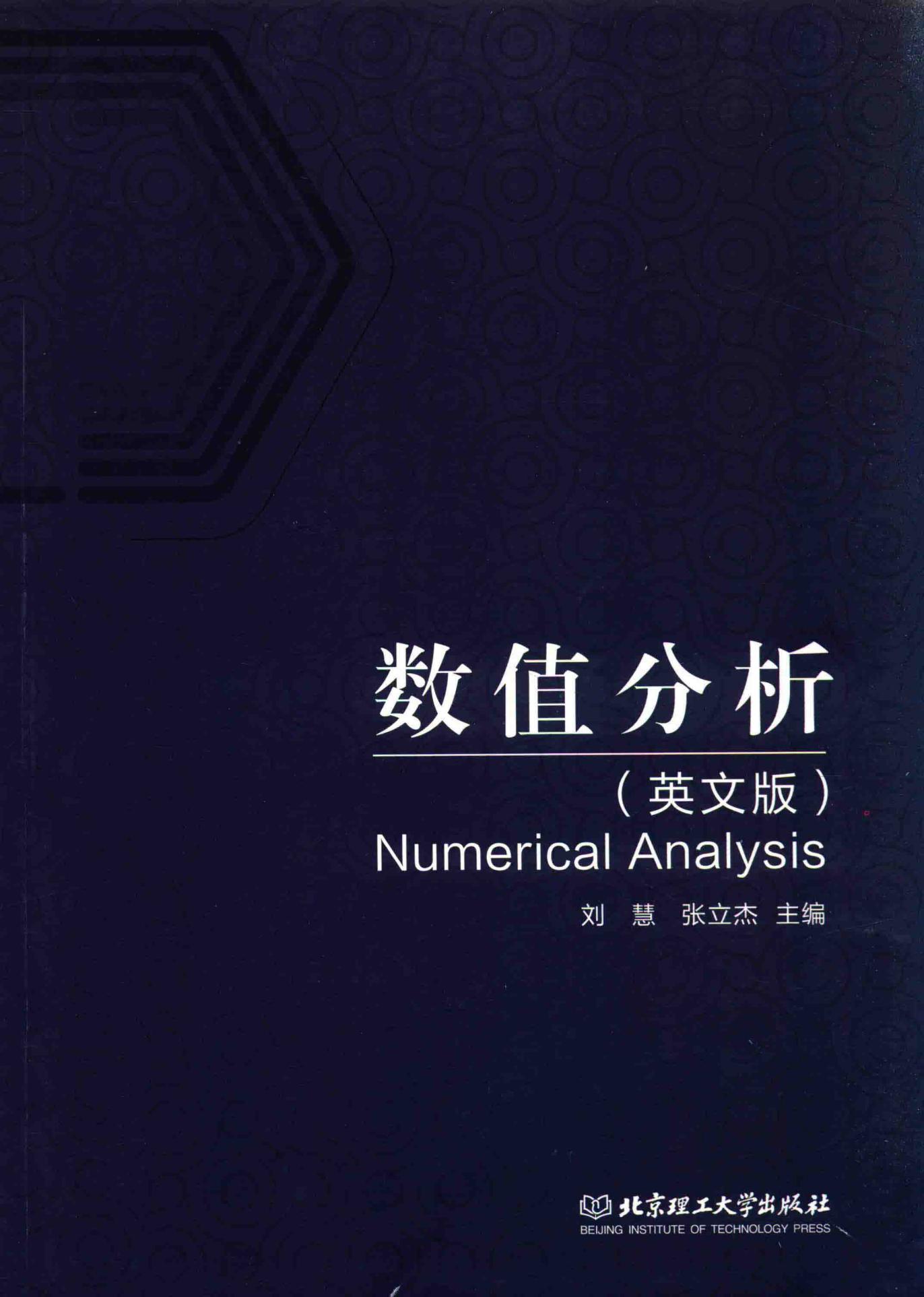




数值分析

(英文版)

Numerical Analysis

刘 慧 张立杰 主编

 北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

数值分析(英文版)

Numerical Analysis

刘 慧 张立杰 主编

 北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

内 容 提 要

本书以英文版的形式介绍面向工程实践和科学计算的各类数学模型的常用数值求解方法以及相关概念和理论,主要内容包括非线性方程数值求解问题、线性方程组数值求解问题、插值与拟合问题、常微分方程初值问题的数值求解问题,以及数值积分和数值微分等。本书对各类模型的解存在理论、数值求解方法的构造原理、数值算法的推演过程、数值方法的收敛性和误差分析等都有较为详细的描述。本书各章配有适量的例题和习题,每一章附有英汉数学词汇表,书末附录一些常用数学表达式的读法,以方便读者对各章节内容的理解、练习和掌握。

本书为适应双语教学特点,专为理工科专业及工程研究生开设的“数值分析”或“计算方法”课程编写的双语教材,适合一学期教学使用。本书可作为理工科大专院校数学、计算机及工程类各专业本科生或研究生的双语课程教材或专业英语教材,也可供从事科学计算的科学工作者做参考。

版权专有 侵权必究

图书在版编目(CIP)数据

数值分析 = Numerical Analysis: 英文/刘慧,张立杰主编. —北京:北京理工大学出版社,2017.9

ISBN 978 - 7 - 5682 - 4847 - 1

I. ①数… II. ①刘…②张… III. ①数值分析 - 英文 IV. ①O241

中国版本图书馆 CIP 数据核字(2017)第 223388 号

出版发行 / 北京理工大学出版社有限责任公司

社 址 / 北京市海淀区中关村南大街 5 号

邮 编 / 100081

电 话 / (010)68914775(总编室)
(010)82562903(教材售后服务热线)
(010)68948351(其他图书服务热线)

网 址 / <http://www.bitpress.com.cn>

经 销 / 全国各地新华书店

印 刷 / 北京国马印刷厂

开 本 / 787 毫米 × 1092 毫米 1/16

印 张 / 9.25

字 数 / 217 千字

版 次 / 2017 年 9 月第 1 版 2017 年 9 月第 1 次印刷

定 价 / 49.00 元

责任编辑 / 梁铜华

文案编辑 / 梁铜华

责任校对 / 周瑞红

责任印制 / 施胜娟

图书出现印装质量问题,请拨打售后服务热线,本社负责调换

前 言

科学计算与理论方法和实验方法并列成为科学研究的基本手段,在现代科学研究和工程实践中的作用毋庸置疑。数值分析,作为科学计算的基础,随着信息化技术应用的广泛深入和计算机技术的推广普及,其作用已渗透到了各学科、各领域的方方面面。目前我国多数理工科高等院校都将“数值分析”或“计算方法”列为理工科专业本科生的必修课程以及工科研究生的学位课程。

教育部在 2001 年制订的《关于加强高等学校本科教学工作,提高教学质量的若干意见》中,明确要求各高校积极开展双语教学,并且建议本科教育要创造条件使用英语等外语进行公共课和专业课教学。青岛理工大学信息与计算科学和数学与应用数学两个专业分别从 2003 年和 2007 年开始进行“数值分析”双语课程的教学改革和实践,先后采用过国外优秀信息类原版教材和国内不同院校改编的英文版教材。但正所谓尺有所短,寸有所长,在双语教学的探索和积累过程中,少有从内容选取、知识推进、方法推演、教学实际等方面都合适的教材。感谢青岛理工大学应用基础型人才培养特色名校建设工程项目的支持,该项目使编者能够有机会鼓足勇气编写《数值分析》英文版教材。同时也感谢北京理工大学出版社的编辑和工作人员,使本书得以顺利出版。

本书面向高等院校设有“数值分析”或“计算方法”课程的各理工类专业,适合作为本科生的双语课程教材或研究生的学位课程教材,亦可作为专业英语教材,还可为从事科学计算的科学工作者提供参考。全书用英文编写,语言力求简练、规范、通俗易懂,既便于教学,也便于阅读。在内容选取上,突出基本概念和方法,强化知识体系的严谨和完整,注重数学思维能力的培养,既便于讲授,也便于自学。

全书共有七章,内容涵盖非线性方程数值求解、线性方程组的直接解法和迭代方法、数值逼近(包括插值、样条和最小二乘拟合)、数值积分方法、微分方程初值问题的数值求解等。全书各章配有适量的例题和习题。为了便于对数值方法的理解和比较,对同一模型不同算法的例题或习题尽可能选择一致,以求在理论分析、算法推演、数值结果和几何直观等方面有更加立体的认知。全书各章配有生词和短语表,汇集该章中涉及的常用数学词汇,为学习习惯也为学习便捷。每一章开篇都有一句谚语,为知识趣味更为激励推动。

限于作者的水平,书中不当之处在所难免,恳请读者给予批评指正(敬请致函 liuhui8259@qut.edu.cn),编者将不胜感激。

刘慧,张立杰
2017 年 4 月

Contents

目录

Chapter 1 Mathematical Review and Introduction

(知识回顾和简介) 1

1.1 Mathematical Review

(知识回顾) 1

1.2 Errors

(误差) 3

1.3 Avoiding the Loss of Accuracy

(避免精度损失) 7

Exercises for Chapter 1 10

Words and Phrases(I) 11

Chapter 2 Solutions of Nonlinear Equations

(非线性方程求解) 12

2.1 Bisection Method

(二分法) 12

2.2 Fixed-Point Iteration

(不动点迭代) 15

2.3 Newton-Raphson Iteration

(牛顿迭代) 20

2.4 Secant Method

(割线法) 25

2.5 Rate of Convergence

(收敛速度) 27

Exercises for Chapter 2 30

Words and Phrases(II) 31

Chapter 3 Solutions of Linear Systems of Equations

(线性方程组求解) 32

3.1 Gaussian Elimination Method

(高斯消去法) 33

3.1.1 Backward Substitution Method

(回代方法) 33

3.1.2	Gaussian Elimination with Backward Substitution (高斯消元回代过程)	35
3.1.3	Gaussian Elimination with Pivoting (列选主元消去法)	38
3.1.4	Gaussian Elimination on Tridiagonal Linear System (三对角线性方程组的高斯消去)	40
3.2	Matrix Factorization (矩阵分解)	43
3.3	Iteration Methods (迭代法)	50
3.3.1	Norms of Vectors and Matrices (向量范数和矩阵范数)	51
3.3.2	Jacobi Iteration (雅克比迭代)	53
3.3.3	Gauss-Seidel Iteration (高斯-塞德尔迭代)	55
3.3.4	Convergence of Iteration Methods (迭代法的收敛性)	56
	Exercises for Chapter 3	61
	Words and Phrases(III)	64
Chapter 4	Interpolation and Polynomial Approximation (插值与多项式逼近)	65
4.1	Interpolation Problem (插值问题)	65
4.2	Lagrange Interpolation (拉格朗日插值)	67
4.3	Newton Interpolation (牛顿插值)	71
4.4	Hermite Interpolation (埃尔米特插值)	75
4.5	Piecewise-Polynomial Interpolation (分段插值)	78
4.6	Cubic Spline Interpolation (三次样条)	80
	Exercises for Chapter 4	86
	Words and Phrases(IV)	87

Chapter 5 Curve Fitting

(曲线拟合)	89
5.1 Least Squares Problem	
(最小二乘问题)	89
5.2 Polynomial Fitting	
(多项式拟合)	90
5.2.1 Least-Squares Line	
(最小二乘直线)	90
5.2.2 Least-Squares Parabola	
(最小二乘抛物线)	92
5.2.3 Polynomial Fitting	
(多项式拟合)	93
5.2.4 Polynomial Wiggle	
(多项式的摆动)	94
5.3 Data Linearization	
(数据线性化)	95
Exercises for Chapter 5	97
Words and Phrases(V)	98

Chapter 6 Numerical Differentiation and Numerical Integration

(数值积分)	99
6.1 Numerical Differentiation	
(数值微分)	99
6.2 Elements of Numerical Quadrature	
(数值积分的基本概念)	102
6.3 Newton-Cotes Integration	
(牛顿 - 科特斯公式)	103
6.4 Composite Numerical Integration	
(复合求积公式)	107
6.5 Romberg Integration	
(龙贝格积分)	110
6.6 Gauss-Legendre Integration	
(高斯 - 勒让德公式)	116
Exercises for Chapter 6	119
Words and Phrases(VI)	122

Chapter 7	Solutions of Ordinary Differential Equations	
	(常微分方程求解)	123
7.1	Elementary Theory of Initial-Value Problems	
	(初值问题基本理论)	124
7.2	Euler's Method	
	(欧拉法)	126
7.3	Modified Euler Method	
	(改进的欧拉法)	128
7.4	Runge-Kutta Methods	
	(龙格-库塔法)	131
	Exercises for Chapter 7	135
	Words and Phrases(VII)	136
References		
	(参考文献)	137
Appendix Math Expressions and Pronunciations		
	(附录 一些数学表达及其读法)	138

A good beginning is half done.

Chapter 1 Mathematical Review and Introduction

This chapter contains a short review of those topics from elementary single-variable calculus that will be needed in later chapters, together with an introduction to error analysis.

1.1 Mathematical Review

Definition 1.1 Let f be a function defined on a set X of real numbers. Then f is said to have the *limit* L at x_0 , written as $\lim_{x \rightarrow x_0} f(x) = L$, if, given any number $\varepsilon > 0$, there exists a number $\delta > 0$, such that $|f(x) - L| < \varepsilon$, whenever $x \in X$ and $0 < |x - x_0| < \delta$.

Definition 1.2 Let f be a function defined on a set X of real numbers and $x_0 \in X$. Then f is *continuous* at x_0 if $\lim_{x \rightarrow x_0} f(x) = f(x_0)$.

Theorem 1.1 (Lagrange's Mean Value Theorem) If f is continuous on the closed interval $[a, b]$, i. e. $f \in C[a, b]$, and f is differentiable on (a, b) , then a number c between a and b exists with

$$f'(c) = \frac{f(b) - f(a)}{b - a}. \quad (1.1)$$

Theorem 1.2 (Rolle's Theorem) Suppose that $f \in C[a, b]$ and f is differentiable on (a, b) . In addition, if $f(a) = f(b)$, then a number c in (a, b) exists with

$$f'(c) = 0. \quad (1.2)$$

Theorem 1.3 (Generalized Rolle's Theorem) Let $x_0, x_1, \dots, x_n$ be $n + 1$ distinct zeros of $f(x)$ in $[a, b]$. Suppose that f and its derivatives of order up to n are all continuous on $[a, b]$. Then there exists a number c in (a, b) such that

$$f^{(n)}(c) = 0. \quad (1.3)$$

Theorem 1.4 (Intermediate Value Theorem) If $f \in C[a, b]$ and u is any number between $f(a)$ and $f(b)$, then there exists a number $c \in (a, b)$ such that

$$f(c) = u. \quad (1.4)$$

Corollary 1.1 (Existence Theorem for Roots) Let $f \in C[a, b]$. If $f(a)$ and $f(b)$ are opposite

in sign, then there exists a number c in (a, b) such that

$$f(c) = 0. \quad (1.5)$$

Theorem 1.5 (Boundedness Theorem) If $f \in C[a, b]$, then there exist numbers M and m such that

$$m \leq f(x) \leq M \quad (1.6)$$

for all $x \in [a, b]$.

Corollary 1.2 Suppose that $f \in C[a, b]$, $x_1, x_2 \in [a, b]$, and c_1, c_2 are positive constants. Then there exists a number c between x_1 and x_2 with the property

$$f(c) = \frac{c_1 f(x_1) + c_2 f(x_2)}{c_1 + c_2}. \quad (1.7)$$

Definition 1.3 The *Riemann integral* of f on $[a, b]$ is the following limit, provided it exists:

$$\int_a^b f(x) dx = \lim_{\max |\Delta x_i| \rightarrow 0} \sum_{i=1}^n f(\xi_i) \Delta x_i, \quad (1.8)$$

where $a = x_0 \leq x_1 \leq \dots \leq x_n = b$, $\Delta x_i = x_i - x_{i-1}$, and ξ_i is arbitrary in $[x_{i-1}, x_i]$ for each $i = 1, 2, \dots, n$.

Theorem 1.6 (Weighted Mean Value Theorem for Integrals) Suppose that $f \in C[a, b]$, the Riemann integral of g exists on $[a, b]$, and $g(x)$ does not change signs on $[a, b]$. Then there exists a number c such that

$$\int_a^b f(x) g(x) dx = f(c) \int_a^b g(x) dx. \quad (1.9)$$

Theorem 1.7 (Taylor's Theorem) Let $f \in C^n[a, b]$ and $x_0 \in [a, b]$. Then for any $x \in [a, b]$, there exists a number $c = c(x)$ between x_0 and x such that

$$f(x) = P_n(x) + R_n(x), \quad (1.10)$$

where, $P_n(x)$, called the n th Taylor polynomial, is in the form

$$P_n(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \dots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n \quad (1.11)$$

and $R_n(x)$, called the error term or the remainder term, in the form

$$R_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!}(x - x_0)^{n+1}. \quad (1.12)$$

1.2 Errors

Numerical answers to problems generally contain errors which arise in two areas—those inherent in mathematical formulation of the problem and those incurred in finding the solution numerically. One source of inherent errors incurs when the mathematical statement of a problem is only an approximation to the physical situation. Such errors, called *modeling errors*, are negligible. Another source of inherent error is the inaccuracies in the physical data. Such errors, called *observational errors*, are also generally negligible when they are caused by inaccuracies in physical constants (e. g. the gravitational constant). Numerical analysis is interested in the errors caused by solving the mathematical models using numerical methods. Generally, two types of errors, called *truncation error* and *round-off error*, are mainly concerned in numerical computing. Truncation error is usually caused by replacement of an infinite (i. e. summation or integration) or infinitesimal (i. e. differentiation) process by a finite approximation.

Example 1.1 The formula

$$f'(x) \approx \frac{f(x+h) - f(x)}{h} \quad (1.13)$$

is not exact for $f'(x)$, but it gives a useful approximation if h is small enough. The formula (1.13) can be obtained by truncating the first two terms of Taylor's series

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2!}f''(x) + \frac{h^3}{3!}f'''(x) + \dots$$

for finite process and the truncation error is

$$\text{err} = \frac{f(x+h) - f(x)}{h} - f'(x) = \frac{h}{2}f''(\xi), \quad (1.14)$$

where ξ is between x and $x+h$.

Table 1.1 lists the results by using formulas (1.13) and (1.14) when we take $f(x) = xe^x$, $x = 1$, and several h values. The exact value of $f'(1)$ is 5.436 563 6...

Table 1.1

h	0.3	0.01	10^{-5}	10^{-8}	10^{-10}
$\widehat{f}'$	6.839 3	5.477 5	5.436 604	5.436 564	5.436 562
$ \text{err} $	1.402 7	0.040 96	4.08×10^{-5}	5.76×10^{-8}	1.35×10^{-6}

Round-off error is produced when a calculator or computer is used to perform real number calculations. It occurs because the arithmetic performed in a machine involves numbers with only a finite number of digits, with the result that the calculations are performed with only approximate representations of the actual numbers.

The machine numbers are usually represented in the *normalized decimal floating-point number* of the form

$$\pm 0. d_1 d_2 \dots d_k \times 10^m, \quad 1 \leq d_1 \leq 9, \quad \text{and} \quad 0 \leq d_i \leq 9 \text{ for } i = 2, 3, \dots, k.$$

A real number within the numerical range of the machine can be normalized in the form

$$y = \pm 0. d_1 d_2 \dots d_k d_{k+1} d_{k+2} \dots \times 10^m.$$

There are two methods to obtain the k th-digit floating-point number of y , denoted $fl(y)$. One method, called *chopping*, is to simply chop off the digits $d_{k+1} d_{k+2} \dots$ to obtain

$$fl(y) = 0. d_1 d_2 \dots d_k \times 10^m.$$

The other method, called *rounding*, adds $5 \times 10^{m-(k+1)}$ to y and then chops the result to obtain a number of the form

$$fl(y) = 0. \delta_1 \delta_2 \dots \delta_k \times 10^m.$$

When rounding, if $d_{k+1} \geq 5$, we add 1 to d_k to obtain $fl(y)$; that is, we are rounding up. If $d_{k+1} < 5$, we merely chop off all but the first k digits; so, rounding down.

Example 1.2 Consider the irrational number $\pi = 3.141\,592\,65\dots$. It is normalized in the decimal floating-point form $\pi = 0.314\,159\,265\dots \times 10^1$.

Then,

$$fl(\pi) = 0.314\,15 \times 10^1 = 3.141\,5 \quad \text{using five-digit chopping,}$$

$$fl(\pi) = (0.314\,15 + 0.000\,01) \times 10^1 = 3.141\,6 \quad \text{using five-digit rounding,}$$

$$\text{and} \quad fl(\pi) = 0.314\,159 \times 10^1 = 3.141\,59 \quad \text{using six-digit rounding.}$$

Using the symbol $\odot$ denotes any one of the arithmetic operations: addition $+$, subtraction $-$, multiplication $\times$, and division $\div$. For any numbers x and y , we will assume a finite-digit arithmetic given by

$$x \odot y = fl(fl(x) \odot fl(y)).$$

For example, using three-digit rounding arithmetic, we have

$$\begin{aligned} \left(\frac{1}{3} - \frac{3}{11}\right) + \frac{3}{20} &= (0.333 - 0.273) + 0.150 \\ &= 0.060\,0 + 0.150 = 0.210. \end{aligned}$$

Definition 1.4 If p^* is an approximation to p , the *absolute error* is

$$e_a = |p - p^*| \quad (1.15)$$

and the *relative error* is

$$e_r = \frac{|p - p^*|}{|p|} \quad (1.16)$$

provided that $p \neq 0$.

Example 1.3 Consider the absolute errors and relative errors in representing p by p^* .

$$(1) \quad p = 0.3000 \times 10^1, \quad p^* = 0.3100 \times 10^1;$$

$$(2) \quad p = 0.3000 \times 10^{-3}, \quad p^* = 0.3100 \times 10^{-3};$$

$$(3) \quad p = 0.3000 \times 10^4, \quad p^* = 0.3100 \times 10^4.$$

Solution Using formula (1.15) and (1.16) respectively, we have the results listed in Table 1.2.

Table 1.2

Exact Value p	0.3000×10^1	0.3000×10^{-3}	0.3000×10^4
Approximation p^*	0.3100×10^1	0.3100×10^{-3}	0.3100×10^4
Absolute Error e_a	0.1	0.1×10^{-4}	0.1×10^3
Relative Error e_r	0.3333×10^1	0.3333×10^1	0.3333×10^1

Example 1.3 indicates that the same relative error, 0.3333×10^1 , occurs for widely varying absolute errors, 0.1, 0.1×10^{-4} , and 100. As the measure of accuracy, the absolute error can be misleading and the relative error is more meaningful since the latter has the consideration on the size of the value itself.

Definition 1.5 $p^* = 0.d_1d_2\dots d_k \times 10^m$ is said to approximate p to n significant digits, or p^* retains n significant digits as an approximation to p , if n is the largest nonnegative integer for which

$$|p - p^*| \leq 0.5 \times 10^{m-n} \quad (1.17)$$

or

$$\frac{|p - p^*|}{|p|} \leq 5 \times 10^{-n}. \quad (1.18)$$

Example 1.4 Assume that $p_1^* = 20.03$, $p_2^* = 20.031$, and $p_3^* = 20.032$ are all approximations to $p = 20.03173$. Determine the number of significant digits for p_1^* , p_2^* and p_3^* , respectively.

Solution Rewriting p , p_1^* , p_2^* , and p_3^* in the floating-point decimal form, respectively, gives

$$p = 0.2003173 \times 10^2,$$

$$p_1^* = 0.2003 \times 10^2,$$

$$p_2^* = 0.20031 \times 10^2,$$

and

$$p_3^* = 0.20032 \times 10^2.$$

Then with $m=2$ and using inequation(1.17), we have

$$|p - p_1^*| = 0.000\,017\,3 \times 10^2 = 0.173 \times 10^{2-4} < 0.5 \times 10^{2-4},$$

$$|p - p_2^*| = 0.000\,007\,3 \times 10^2 = 0.073 \times 10^{2-4} < 0.5 \times 10^{2-4},$$

and

$$|p - p_3^*| = 0.000\,002\,7 \times 10^2 = 0.27 \times 10^{2-5} \leq 0.5 \times 10^{2-5}.$$

Thus, p_1^* , p_2^* , and p_3^* agree with p to 4, 4, and 5 significant digits, respectively.

Theorem 1.8 (a) If $p^* = \pm 0.d_1 d_2 \dots d_n \times 10^m$ is an approximation to p accurate to n significant digits, then the bound of $\frac{|p - p^*|}{|p|}$ is $\frac{1}{2d_1} \times 10^{-n+1}$.

(b) If the bound of $\frac{|p - p^*|}{|p|}$ is $\frac{1}{2(d_1 + 1)} \times 10^{-n+1}$, then $p^* = \pm 0.d_1 d_2 \dots d_n \times 10^m$ has n significant digits.

Proof (a) The hypothesis, $p^* = \pm 0.d_1 d_2 \dots d_n \times 10^m$ has n significant digits, means that

$$|p - p^*| \leq \frac{1}{2} \times 10^{m-n}.$$

Therefore,

$$\frac{|p - p^*|}{|p|} \leq \frac{\frac{1}{2} \times 10^{m-n}}{d_1 \times 10^{m-1}} = \frac{1}{2d_1} \times 10^{-n+1}.$$

(b) $p^* = \pm 0.d_1 d_2 \dots d_n \times 10^m$ has n significant digits since

$$\begin{aligned} |p - p^*| &= \frac{|p - p^*|}{|p|} \cdot |p| \\ &\leq \frac{1}{2(d_1 + 1)} \times 10^{-n+1} \cdot (d_1 + 1) \times 10^{m-1} \\ &= \frac{1}{2} \times 10^{m-n}. \end{aligned}$$

Example 1.5 Let p^* be an approximation to $\sqrt{70}$. Determine the least number of significant digits of p^* such that the relative error bound is within 10^{-3} .

Solution Note that the first nonzero digit of $\sqrt{70}$ is $d_1 = 8$. Using Theorem 1.7(a), we have

$$\frac{|p - p^*|}{|p|} \leq \frac{1}{2d_1} \times 10^{-n+1} = \frac{1}{16} \times 10^{-n+1} \leq 10^{-3}$$

and $n \geq 2.7959$. Thus, p^* should have at least 3 significant digits to guarantee the given accuracy.

1.3 Avoiding the Loss of Accuracy

In the process of computation, errors will be accumulated after initial errors are given. Several rules in computing should be followed in order to keep results as accurate as possible.

Rule I Avoid subtraction of nearly equal numbers

For example, using the quadratic formula and four-digit rounding arithmetic to find the roots of $x^2 + 62.10x + 1 = 0$ produces

$$x_1^* = \frac{-62.10 + \sqrt{(62.10)^2 - 4.000 \times 1.000 \times 1.000}}{2.000 \times 1.000}$$

and

$$x_2^* = \frac{-62.10 - \sqrt{(62.10)^2 - 4.000 \times 1.000 \times 1.000}}{2.000 \times 1.000}.$$

Calculating directly gives

$$x_1^* = \frac{-62.10 + \sqrt{3852}}{2.000} \approx \frac{-62.10 + 62.06}{2.000} = \frac{-0.04000}{2.000} = -0.02000$$

and

$$x_2^* \approx \frac{-62.10 - 62.06}{2.000} = \frac{-124.2}{2.000} = -62.10$$

while the actual roots are approximately $x_1 = -0.01610723$ and $x_2 = -62.08390$.

x_1^* , a poor approximation with a large relative error

$$\frac{|-0.01611 - (-0.02000)|}{|-0.01611|} \approx 2.4 \times 10^{-1}$$

is obtained by the subtraction of nearly equal numbers

$$-62.10 + \sqrt{3852} = -62.10 + 62.06 = -0.0400.$$

An alternative method for more accurate x_1^* is to rationalize the numerator:

$$\begin{aligned} x_1^* &= \frac{-62.10 + \sqrt{(62.10)^2 - 4.000}}{2.000} \\ &= \frac{(-62.10)^2 - ((62.10)^2 - 4.000)}{2.000 \times (-62.10 - \sqrt{(62.10)^2 - 4.000})} \end{aligned}$$

$$= \frac{2.000}{-62.10 - \sqrt{(62.10)^2 - 4.000}}.$$

Then, $x_1^* = -0.01610$ with a small relative error 6.2×10^{-4} .

Reforming the calculations is a useful technique to avoid the subtraction of nearly equal numbers.

Rule II Avoid division by numbers with small magnitude

For example, the solutions of the linear equations

$$E_1: 0.003000x_1 + 59.14x_2 = 59.17,$$

$$E_2: 5.291x_1 - 6.130x_2 = 46.78$$

are $x_1 = 10.00$ and $x_2 = 1.000$. Now we find them using elimination approach and the four-digit rounding arithmetic.

Let

$$m_{21} = \frac{5.291}{0.003000} = 1763.66\ldots \approx 1764.$$

Performing the elementary transformations $(E_2 - m_{21}E_1) \rightarrow E_2$ gives

$$0.003000x_1 + 59.14x_2 = 59.17,$$

$$-104300x_2 = -104400,$$

and then using back substitutions gives

$$x_2^* = 1.001 \text{ and } x_1^* = \frac{59.17 - 59.14 \times 1.001}{0.003000} = -10.00.$$

What a poor approximation of x_1^* ! The reason is the division by a small number 0.003000.

Change the order of Eq. E_1 and E_2 :

$$5.291x_1 - 6.130x_2 = 46.78,$$

$$0.003000x_1 + 59.14x_2 = 59.17.$$

Repeating the elimination process gives the accurate values $x_1 = 10.00$ and $x_2 = 1.000$.

This technique, called pivoting for solving linear systems of equations, will be discussed in Chapter 3.

Rule III Avoid big number “swallowing” small numbers

For example, calculate the sum $259\,453 + \sum_{k=1}^{100} \delta_k$ where $\delta_k = 0.4$.

If addition operations are performed successively using six-digit rounding arithmetic, then

$$\begin{aligned} 259\,453 + \sum_{k=1}^{1\,000} \delta_k &= 259\,453 + 0.4 + \sum_{k=1}^{999} \delta_k \\ &\approx 0.259\,453 \times 10^6 + 0.000\,000 \times 10^6 + \sum_{k=1}^{999} \delta_k \\ &\approx 0.259\,453 \times 10^6 + \sum_{k=1}^{998} \delta_k \approx \cdots \approx 259\,453. \end{aligned}$$

The result indicates that one by one, 100 small numbers are “swallowed” by the large number 259 453. In order to avoid this, we mix all small numbers together firstly,

$$\sum_{k=1}^{1\,000} \delta_k = 0.4 \times 10^3,$$

and then add the sum above in the large number:

$$\sum_{k=1}^{1\,000} \delta_k + 259\,453 = 0.000\,400 \times 10^6 + 0.259\,453 \times 10^6 = 259\,853.$$

Rule IV Reduce the number of calculations

For example, the exact value of $f(x) = x^3 - 6.1x^2 + 3.2x + 1.5$ at $x = 4.71$ is $-14.263\,899$. By using three-digit rounding arithmetic to evaluate $f(4.71)$, we have

$$\begin{aligned} f(4.71) &= (4.71)^3 - 6.1 \times (4.71)^2 + 3.2 \times 4.71 + 1.5 \\ &\approx 105 - 6.1 \times 22.2 + 3.2 \times 4.71 + 1.5 \approx 105 - 135 + 15.1 + 1.5 \\ &= ((105 - 135) + 15.1) + 1.5 = -13.4 \end{aligned} \tag{1.19}$$

with the relative error 0.06. On the other hand, rewrite $f(x)$ in the *nested form*

$$f(x) = ((x - 6.1)x + 3.2)x + 1.5$$

and we have

$$f(4.71) = ((4.71 - 6.1) \times 4.71 + 3.2) \times 4.71 + 1.5 = -14.3 \tag{1.20}$$

with the relative error 0.002 5. In fact, the nested form reduces five multiplications and three additions in Eq. (1.19) to three multiplications and three additions in Eq. (1.20). The nesting approach is also called Horner’s method and often applied to evaluate polynomials.