NUMERICAL METHODS WITH APPLICATIONS

(Abridged Edition)

Authors

AUTAR KAW, University of South Florida

http://www.autarkaw.com

E. ERIC KALU, Florida A&M University

http://www.eng.fsu.edu/~ekalu/

Contributors

GLEN BESTERFIELD, University of South Florida

SUDEEP SARKAR, University of South Florida

HENRY WELCH, Milwaukee School of Engineering

ALI YALCIN, University of South Florida

There is nothing noble about being superior to another man; the true nobility lies in being superior to your previous self - Upanishads.

NUMERICAL METHODS WITH APPLICATIONS

(Abridged Edition)

First Edition

AUTAR KAW
University of South Florida
E. ERIC KALU
Florida A&M University

To Sherrie, Candace and Angelie
AKK

To Ngozi, Ola, Erinma and Egwuchukwu
EEK

AUTAR K KAW

Autar K. Kaw is a Professor of Mechanical Engineering at the University of South Florida, Tampa. Professor Kaw obtained his B.E. (Hons.) degree in Mechanical Engineering from Birla Institute of Technology and Science, India in 1981. He received his Ph.D. degree in 1987 and M.S. degree in 1984, both in Engineering Mechanics from Clemson University, SC. He joined the faculty of University of South Florida, Tampa in 1987. He has also been a Maintenance Engineer (1982) for Ford-Escorts Tractors, India, and a Summer Faculty Fellow (1992) and Visiting Scientist (1991) at Wright Patterson Air Force Base.

Professor Kaw's main scholarly interests are in engineering education research, bridge design, thermal stresses, engineering software, computational nanomechanics, and fracture. His research has been funded by National Science Foundation, Air Force Office of Scientific Research, Florida Department of Transportation, Research and Development Laboratories, Systran, Wright Patterson Air Force Base, and Montgomery Tank Lines.

Professor Kaw is a Fellow of the American Society of Mechanical Engineers (ASME) and a member of the American Society of Engineering Education (ASEE). He has written more than forty journal papers and developed several software instructional programs for courses such as Mechanics of Composites and Numerical Methods.

Professor Kaw has published textbooks on *Mechanics of Composite Materials* (CRC press), and *Introduction to Matrix Algebra* (autarkaw.com).

Professor Kaw received the Florida Professor of the Year Award from the Council for Advancement and Support of Education (CASE) and Carnegie Foundation for Advancement of Teaching (CFAT) in 2004, American Society of Mechanical Engineers (ASME) Curriculum Innovation Award in 2004, Archie Higdon Mechanics Educator Award from the American Society of Engineering Education (ASEE) in 2003, State of Florida Teaching Incentive Program Award in 1994 and 1997, American Society of Engineering Education (ASEE) New Mechanics Educator Award in 1992, and the Society of Automotive Engineers (SAE) Ralph Teetor Award in 1991.

E. ERIC KALU

E. Eric. Kalu is an Associate Professor of Chemical & Biomedical Engineering at the Florida A&M University-Florida State University College of Engineering, Tallahassee. Dr Kalu received a B.Sc (Hons) degree with First Class in Chemical Engineering from University of Lagos, Nigeria in 1984 and in 1988 was awarded a MASc (Chemical Engineering) from the University of British Columbia, Canada. He also received a Ph.D. degree from Texas A&M University, TX in Chemical Engineering in 1991. Before joining the faculty at Florida A&M University in 1995, he worked as a Senior Research Engineer (1991 – 1993) for Monsanto Chemical Company in St Louis, MO and Research Assistant Professor at University of South Carolina, Columbia (1994). He has also been a Summer Faculty Fellow (2007) at Sandia National Laboratories.

Dr Kalu's research interests are in electrochemical & nanomaterials engineering for sustainable energy and environmental systems, engineering education and mentoring. Mathematical modeling of electrochemical and thermochemical systems research are areas of key interest. Dr Kalu has written several journal papers in his areas of interest.

He is a member of the American Institute of Chemical Engineers (AIChE) and a member of the Electrochemical Society (ECS).

Dr Kalu received the NASA Faculty Fellowship Award (2004), Lockheed Martin E&M Minority Institution Award (1998) and High Merit Award for Pioneering Nanotechnology Research Masscal Scientific Instruments (2007). He was also awarded Florida State University First year Assistant Professor Research Award in 1996.

GLEN BESTERFIELD

Glen Besterfield has held the position of Associate Dean for Undergraduate Studies and Student Academic Success at the University of South Florida, Tampa since 2005. As Associate Dean, he is tasked with all aspects of student success at USF from matriculation through graduation. Responsibilities include the orientation process, academic advising, career services, class scheduling, tracking the academic progress of students, first generation access and success, student tutorials and learning services, and academics of student athletes, to name a few.

Dr. Besterfield has been a member of the Department of Mechanical Engineering Faculty at USF since 1988. He received his BS degree in Mechanical Engineering from the University of Missouri, Rolla in 1982, MS degree from Purdue University in 1984, and his Ph.D. from Northwestern University in 1989. Prior to joining the faculty in Mechanical Engineering, he worked in various capacities at Argonne National Laboratory, IBM Corporation, and McDonnell Douglas Corporation.

Dr. Besterfield's primary areas of research and scholarly work are in the fields of student success, total quality management, bridge design, thermal stresses, computational mechanics, and rehabilitation engineering. As a result of his efforts in these areas, Dr. Besterfield has received national recognition in several areas of great importance to society. His research has been funded by National Science Foundation, Air Force Office of Scientific Research, Florida Department of Transportation, US Department of Veterans Affairs, Florida Department of Labor and Employment Security – Division of Vocational Rehabilitation, Wolff Controls, and Montgomery Tank Lines. Professor Besterfield has also co-written a book entitled *Total Quality Management* published by Prentice-Hall.

SUDEEP SARKAR

Sudeep Sarkar received his B.Tech degree in Electrical Engineering from the Indian Institute of Technology, Kanpur, in 1988. He received the M.S. and Ph.D. degrees in Electrical Engineering, on a University Presidential Fellowship, from The Ohio State University, Columbus, in 1990 and 1993, respectively. Since 1993, he has been with the Computer Science and Engineering Department at the University of South Florida, Tampa, where he is currently a Professor. His research interests include perceptual organization, automated American Sign Language recognition, biometrics, gait recognition, and nano-computing. He is the co-author of the book *Computing Perceptual Organization in Computer Vision* published by World Scientific. He is also the co-editor of the book *Perceptual Organization for Artificial Vision Systems* published by Kluwer Publishers.

He is the recipient of the National Science Foundation CAREER award in 1994, the USF Teaching Incentive Program Award for undergraduate teaching excellence in 1997, the Outstanding Undergraduate Teaching Award in 1998, and the Theodore and Venette Askounes-Ashford Distinguished Scholar Award in 2004. He served on the editorial boards for the IEEE Transactions on Pattern Analysis and Machine Intelligence (1999-2003) and Pattern Analysis & Applications Journal during 2000-01. He is currently serving on the editorial boards of the Pattern Recognition Journal, IET Computer Vision, Image and Vision Computer, and the IEEE Transactions on Systems, Man, and Cybernetics, Part-B. He is a Fellow of the International Association of Pattern Recognition.

ALI YALCIN

Prof. Ali Yalcin received his B.S., M.S., and Ph.D. degrees in Industrial and Systems Engineering from Rutgers University, New Brunswick New Jersey in 1995, 1997 and 2000. He is currently an Associate Professor at the University of South Florida, Industrial and Management Systems Engineering Department, and an Associate Faculty member of the Center for Urban Transportation Research.

His research interests include modeling, analysis, and control of discrete event systems, production planning and control, industrial information systems, data analysis and knowledge discovery, and engineering education research. He has taught courses in the areas of systems modeling and analysis, information systems design, production planning, facilities design, and systems simulation. He has publications in the areas of control of automated manufacturing systems, transportation systems, and autonomous vehicles, business process modeling, freight transportation systems analysis, people logistics, manufacturing and service information systems, and engineering education research. He co-authored the textbook *Design of Industrial Information Systems* published by Elsevier. The textbook received the 2007 IIE/Joint Publishers Book-of-the-Year Award.

PREFACE

This textbook is written for an undergraduate course in numerical methods for engineering and science. The book has been written in a holistic manner so that the student gets a strong fundamental knowledge of various numerical methods and their applications. Combined with the freely accessible web-based resources at http://numericalmethods.eng.usf.edu, the student can adapt their learning styles and preferences to learn numerical methods.

In 2002, National Science Foundation funded a prototype proposal on Holistic Numerical Methods to develop various resources for typical numerical methods topics of interpolation and solution of nonlinear equations. With the success of this proposal, NSF continued to fund the proposal for other topics of numerical methods via two more multi-university grants in 2004-07 and 2008-10. This funding has so far resulted in complete resources on a typical Numerical Methods course. These resources include textbook chapters, PowerPoint presentations, worksheets in MATLAB, MATHEMATICA, Maple and MathCAD, multiple-choice tests, experiments, video lectures, and a blog (http://autarkaw.wordpress.com). All these resources are available at http://numericalmethods.eng.usf.edu.

The book is abridged because of the following reasons:

1) It is being written primarily for Arizona State University and University of South Florida and follows their respective complete syllabi.
2) The book needs to be kept under the 740-page limit set by the publisher for perfect-bound binding.
3) We believe in keeping textbook prices low.

However, the abridged nature does not sacrifice the level of content available. The chapters that are not in the printed book can be viewed at http://numericalmethods.eng.usf.edu/topics/textbook_index.html.

If you are an instructor of a numerical methods course and you are interested in adopting the book and want to customize it based on your syllabus, please contact the first author at autarkaw@yahoo.com.

We continue to provide the resources free of charge while selling the printed book to gain sustainability as required by our sponsor. *Items available in the book that are not on the web are exercise sets, table of contents, and an index.*

The book is divided into eight topics:

1. Introduction to Scientific Computing,
2. Differentiation,
3. Nonlinear Equations,
4. Simultaneous Linear Equations,
5. Interpolation,
6. Regression,
7. Integration, and
8. Ordinary Differential Equations.

Each topic is covered in several separate chapters because we intend to keep the chapters short and independent. This allowed us to customize the book based on your needs. Supplemental material is always available from the website. Just go to the main website

http://numericalmethods.eng.usf.edu and click on the numerical method of your choice. You will have access to PowerPoint presentations, worksheets, and a multiple-choice test. By June 2009, we will be adding broadcast quality (also available via YouTube at http://www.youtube.com/kawautar) instructional audiovisual content for each numerical method. By December 2010, we will have additional topics of Optimization, Partial Differential Equations, and Fast Fourier Transforms available.

The chapters in the book are numbered as Chapter XX.YY. The XX stands for the topic number while YY is the chapter number within that topic. Most of the chapters are followed by a multiple-choice test based on Bloom's taxonomy (you can take the quizzes online also at http://numericalmethods.eng.usf.edu/assessment_text.html) and a problem set.

Chapter 01.YY introduces scientific computing by taking a real-life example to show that solving an engineering problem requires one to develop a mathematical model, solve the model, and then implement the corresponding solution. This is followed by discussion of sources of numerical error and their measurement, binary and floating-point representation of numbers, and propagation of errors.

Chapters 02.YY through **Chapter 08.YY** cover Differentiation, Nonlinear Equations, Simultaneous Linear Equations, Interpolation, Regression, Integration, and Ordinary Differential Equations, respectively. Each of these topics start with an example of application of the mathematical procedure (e.g. differentiation, nonlinear equations, etc) from each of the seven engineering majors. Background information needed to understand the numerical method is also given. For example, for differentiation, a primer chapter (available online) has been written to review the background from the differential calculus course; for nonlinear equations, quadratic equations are reviewed from the college algebra course; for integration, a primer chapter (available online) has been written to review the background from the integral calculus course. Then numerical methods used to solve the mathematical procedure are shown complete with examples from general engineering. If you want to see how a numerical method works with examples from a different engineering major of your choice, go to http://numericalmethods.eng.usf.edu/numerical_methods_topic_major_language.html. Each chapter is followed by a multiple-choice test and a problem set.

We would like to thank - Sri Harsha Garapati, Luke Snyder, Eric Marvella, Sue Britten, and Matthew Emmons for reformatting and typing the textbook. Sean Rodby's painstaking proofreading has been critical in maintaining accuracy of the contents of the book. We would like to thank Cuong Nguyen, Praveen Chalasani, Michael Keteltas, and Luke Snyder for contributions to the book. Kaw would like to thank his spouse, Sherrie, and children Candace and Angelie, who encouraged him to co-write this textbook.

We would like to thank Professors Melvin Corley of Louisiana Technical University, Duc Nguyen of Old Dominion University, Tianxia Zhao of Indiana University-Purdue University, Fort Wayne, and Xudong Jia of California State Polytechnic University for reviewing the contents of the website which included the textbook notes that are in this book.

The first author has written the major portion of the material in this book. The second author has written most of the chapters on regression. The four contributors to this book have written the real-life problems from their engineering majors of expertise. All the authors and contributors are acknowledged at the end of each chapter.

For further information, please visit the book website at http://autarkaw.com/books/numericalmethods/index.html. There you will find links to the additional resources on each numerical method topic and answers to selected problems.

We would appreciate feedback, questions, or comments that you may have on the book or the numerical methods project. You can contact the first author, Autar Kaw, via

Email: autarkaw@eng.usf.edu
URL: http://autarkaw.com
Tel: 813.974.5626
Mailing Address: Department of Mechanical Engineering Department, University of South Florida, 4202 East Fowler Avenue, ENB118, Tampa, FL 33620-5350.

TABLE OF CONTENTS

Chapter 01.01
Introduction to Numerical Methods

After reading this chapter, you should be able to:

1. *understand the need for numerical methods, and*
2. *go through the stages (mathematical modeling, solving and implementation) of solving a particular physical problem.*

Mathematical models are an integral part in solving engineering problems. Many times, these mathematical models are derived from engineering and science principles, while at other times the models may be obtained from experimental data.

Mathematical models generally result in need of using mathematical procedures that include but are not limited to

(A) differentiation,
(B) nonlinear equations,
(C) simultaneous linear equations,
(D) curve fitting by interpolation or regression,
(E) integration, and
(F) differential equations.

These mathematical procedures may be suitable to be solved exactly as you must have experienced in the series of calculus courses you have taken, but in most cases, the procedures need to be solved approximately using numerical methods. Let us see an example of such a need from a real-life physical problem.

To make the fulcrum (Figure 1) of a bascule bridge, a long hollow steel shaft called the trunnion is shrink fit into a steel hub. The resulting steel trunnion-hub assembly is then shrink fit into the girder of the bridge.

Figure 1 Trunnion-Hub-Girder (THG) assembly.

This is done by first immersing the trunnion in a cold medium such as a dry-ice/alcohol mixture. After the trunnion reaches the steady state temperature of the cold medium, the trunnion outer diameter contracts. The trunnion is taken out of the medium and slid through the hole of the hub (Figure 2).

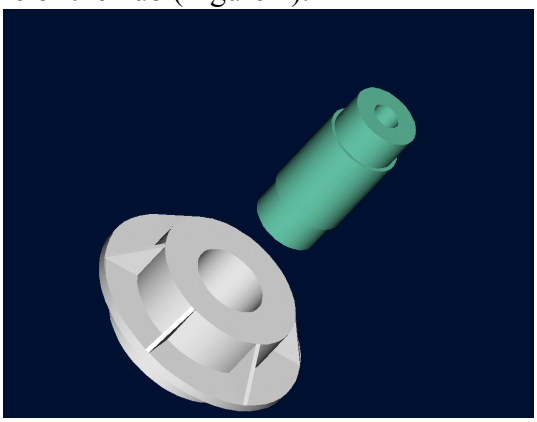


Figure 2 Trunnion slided through the hub after contracting

When the trunnion heats up, it expands and creates an interference fit with the hub. In 1995, on one of the bridges in Florida, this assembly procedure did not work as designed. Before the trunnion could be inserted fully into the hub, the trunnion got stuck. Luckily, the trunnion was taken out before it got stuck permanently. Otherwise, a new trunnion and hub would needed to be ordered at a cost of $50,000. Coupled with construction delays, the total loss could have been more than a hundred thousand dollars.

Why did the trunnion get stuck? This was because the trunnion had not contracted enough to slide through the hole. Can you find out why?

A hollow trunnion of outside diameter $12.363"$ is to be fitted in a hub of inner diameter $12.358"$. The trunnion was put in dry ice/alcohol mixture (temperature of the fluid - dry ice/alcohol mixture is $-108°\text{F}$) to contract the trunnion so that it can be slid through the hole of the hub. To slide the trunnion without sticking, a diametrical clearance of at least $0.01"$ is required between the trunnion and the hub. Assuming the room temperature is $80°\text{F}$, is immersing the trunnion in dry-ice/alcohol mixture a correct decision?

To calculate the contraction in the diameter of the trunnion, the thermal expansion coefficient at room temperature is used. In that case the reduction ΔD in the outer diameter of the trunnion is

$$\Delta D = D\alpha\Delta T \qquad (1)$$

where

$D =$ outer diameter of the trunnion,

$\alpha =$ coefficient of thermal expansion coefficient at room temperature, and

$\Delta T =$ change in temperature,

Given

$$D = 12.363"$$

$$\alpha = 6.47\times10^{-6}\text{in/in/°F at } 80°\text{F}$$

$$\Delta T = T_{fluid} - T_{room}$$

$$= -108 - 80$$

$= -188\,°\text{F}$

where

T_{fluid} = temperature of dry-ice/alcohol mixture

T_{room} = room temperature

the reduction in the outer diameter of the trunnion is given by

$$\Delta D = (12.363)\left(6.47\times10^{-6}\right)(-188)$$
$$= -0.01504"$$

So the trunnion is predicted to reduce in diameter by $0.01504"$. But, is this enough reduction in diameter? As per specifications, the trunnion needs to contract by

= trunnion outside diameter - hub inner diameter + diametric clearance

= 12.363 – 12.358 + 0.01

$= 0.015"$

So according to his calculations, immersing the steel trunnion in dry-ice/alcohol mixture gives the desired contraction of greater than $0.015"$ as the predicted contraction is $0.01504"$. But, when the steel trunnion was put in the hub, it got stuck. Why did this happen? Was our mathematical model adequate for this problem or did we create a mathematical error?

As shown in Figure 3 and Table 1, the thermal expansion coefficient of steel decreases with temperature and is not constant over the range of temperature the trunnion goes through. Hence, Equation (1) would overestimate the thermal contraction.

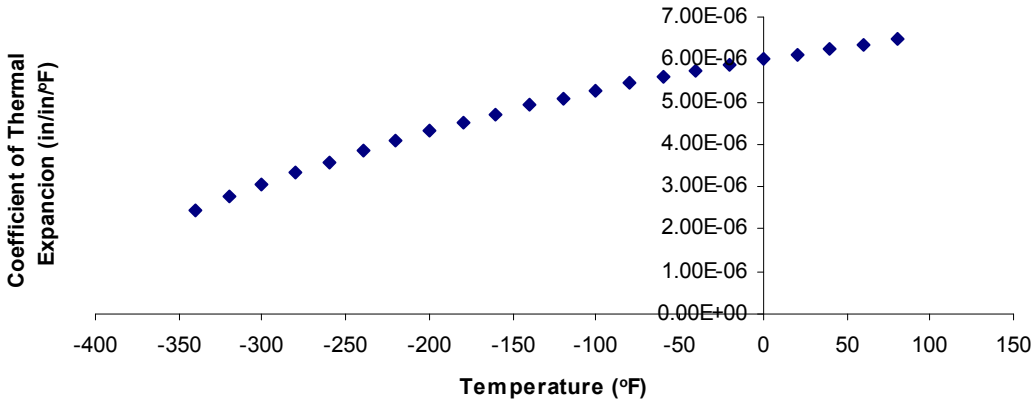


Figure 3 Varying thermal expansion coefficient as a function of temperature for cast steel.

The contraction in the diameter of the trunnion for which the thermal expansion coefficient varies as a function of temperature is given by

$$\Delta D = D\int_{T_{room}}^{T_{fluid}} \alpha dT \tag{2}$$

So one needs to curve fit the data to find the coefficient of thermal expansion as a function of temperature. This is done by regression where we best fit a curve through the data given in Table 1. In this case, we may fit a second order polynomial

$$\alpha = a_0 + a_1 \times T + a_2 \times T^2 \tag{3}$$

Table 1 Instantaneous thermal expansion coefficient as a function of temperature.

Temperature	Instantaneous Thermal Expansion
°F	μin/in/°F
80	6.47
60	6.36
40	6.24
20	6.12
0	6.00
-20	5.86
-40	5.72
-60	5.58
-80	5.43
-100	5.28
-120	5.09
-140	4.91
-160	4.72
-180	4.52
-200	4.30
-220	4.08
-240	3.83
-260	3.58
-280	3.33
-300	3.07
-320	2.76
-340	2.45

The values of the coefficients in the above Equation (3) will be found by polynomial regression (we will learn how to do this later in Chapter 06.04). At this point we are just going to give you these values and they are

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 6.0217\times10^{-6} \\ 6.2782\times10^{-9} \\ -1.2218\times10^{-11} \end{bmatrix}$$

to give the polynomial regression model (Figure 4) as

$$\alpha = a_0 + a_1 T + a_2 T^2$$
$$= 6.0217\times10^{-6} + 6.2782\times10^{-9}\,\mathrm{T} - 1.2218\times10^{-11}\,\mathrm{T}^2$$

Knowing the values of a_0, a_1 nd a_2, we can then find the contraction in the trunnion diameter as

$$\Delta D = D\int_{T_{room}}^{T_{fluid}} (a_0 + a_1 T + a_2 T^2)dT$$
$$= D[a_0(T_{fluid} - T_{room}) + a_1\frac{({T_{fluid}}^2 - {T_{room}}^2)}{2} + a_2\frac{({T_{fluid}}^3 - {T_{room}}^3)}{3}] \quad (4)$$

which gives

$$\Delta D = 12.363\left[\begin{array}{l} 6.0217\times10^{-6}\times(-108-80)+6.2782\times10^{-9}\dfrac{\left((-108)^2-(80)^2\right)}{2} \\ \qquad -1.2118\times10^{-12}\dfrac{((-108)^3-(80)^3)}{3} \end{array}\right]$$

$$= 0.013783"$$

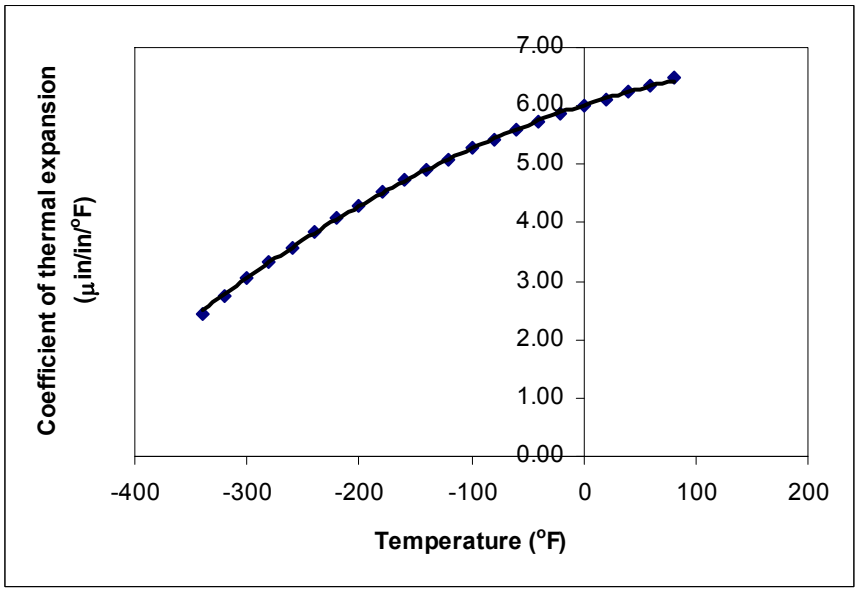


Figure 4 Second order polynomial regression model for coefficient of thermal expansion as a function of temperature.

What do we find here? The contraction in the trunnion is not enough to meet the required specification of $0.015"$.

So here are some questions that you may want to ask yourself?

1. What if the trunnion were immersed in liquid nitrogen (boiling temperature $= -321\,°F$)? Will that cause enough contraction in the trunnion?
2. Rather than regressing the thermal expansion coefficient data to a second order polynomial so that one can find the contraction in the trunnion OD, how would you use Trapezoidal rule of integration for unequal segments? What is the relative difference between the two results?
3. We chose a second order polynomial for regression. Would a different order polynomial be a better choice for regression? Is there an optimum order of polynomial you can find?

As mentioned at the beginning of this chapter, we generally see mathematical procedures that require the solution of nonlinear equations, differentiation, solution of simultaneous linear equations, interpolation, regression, integration, and differential equations. A physical example to illustrate the need for each of these mathematical procedures is given in the beginning of each chapter. You may want to look at them now to understand better why we need numerical methods in everyday life.

INTRODUCTION, APPROXIMATION AND ERRORS	
Topic	Introduction to Numerical Methods
Summary	Textbook notes of Introduction to Numerical Methods
Major	General Engineering
Authors	Autar Kaw
Date	December 7, 2008
Web Site	http://numericalmethods.eng.usf.edu

Multiple-Choice Test

Chapter 01.01
Introduction to Numerical Methods

1. Solving an engineering problem requires four steps. In order of sequence, the four steps are
   (A) formulate, solve, interpret, implement
   (B) solve, formulate, interpret, implement
   (C) formulate, solve, implement, interpret
   (D) formulate, implement, solve, interpret

2. One of the roots of the equation $x^3 - 3x^2 + x - 3 = 0$ is
   (A) -1
   (B) 1
   (C) 3
   (D) $\sqrt{3}$

3. The solution to the set of equations
   $$25a + b + c = 25$$
   $$64a + 8b + c = 71$$
   $$144a + 12b + c = 155$$
   most nearly is $(a,b,c) =$
   (A) (1,1,1)
   (B) (1,-1,1)
   (C) (1,1,-1)
   (D) does not have a unique solution.

4. The exact integral of
   $$\int_0^{\frac{\pi}{4}} 2\cos 2x dx$$
   is most nearly
   (A) -1.000
   (B) 1.000
   (C) 0.000
   (D) 2.000

5. The value of $\frac{dy}{dx}(1.0)$, given $y = 2\sin(3x)$ most nearly is

(A) -5.9399
(B) -1.980
(C) 0.31402
(D) 5.99178

6. The form of the exact solution of the ordinary differential equation $2\frac{dy}{dx} + 3y = 5e^{-x}$, $y(0) = 5$ is

(A) $Ae^{-1.5x} + Be^{x}$
(B) $Ae^{-1.5x} + Be^{-x}$
(C) $Ae^{1.5x} + Be^{-x}$
(D) $Ae^{-1.5x} + Bxe^{-x}$

Answers

1. A
2. C
3. C
4. B
5. A
6. B

Problem Set

Chapter 01.01
Introduction to Numerical Methods

1. Give one example of an engineering problem where each of the following mathematical procedure is used. If possible, draw from your experience in other classes or from any professional experience you have gathered to date.
   a) Differentiation
   b) Nonlinear equations
   c) Simultaneous linear equations
   d) Regression
   e) Interpolation
   f) Integration
   g) Ordinary differential equations

2. Only using your nonprogrammable calculator, find the root of
   $$x^3 - 0.165x^2 + 3.993 \times 10^{-4} = 0$$
   by any method (show your work).

3. Solve the following system of simultaneous linear equations by any method
   $$25a + 5b + c = 106.8$$
   $$64a + 8b + c = 177.2$$
   $$144a + 12b + c = 279.2$$

4. You are given data for the upward velocity of a rocket as a function of time in the table below.

t	$v(t)$
s	m/s
10	305.31
20	701.53

   Find the velocity at $t = 16\text{s}$.

5. Integrate exactly
   $$\int_0^{\pi/2} \sin 2x \, dx$$

6. Find
   $$\frac{dy}{dx}(x = 1.4)$$
   given
   $$y = e^x + \sin(x)$$

7. Solve the following ordinary differential equation exactly

$$\frac{dy}{dx} + y = e^{-x}, y(0) = 5$$

Also find $y(0)$, $\frac{dy}{dx}(0)$, $y(2.5)$, $\frac{dy}{dx}(2.5)$

Chapter 01.02
Measuring Errors

After reading this chapter, you should be able to:

1. *find the true and relative true error,*
2. *find the approximate and relative approximate error,*
3. *relate the absolute relative approximate error to the number of significant digits at least correct in your answers, and*
4. *know the concept of significant digits.*

In any numerical analysis, errors will arise during the calculations. To be able to deal with the issue of errors, we need to

(A) identify where the error is coming from, followed by
(B) quantifying the error, and lastly
(C) minimize the error as per our needs.

In this chapter, we will concentrate on item (B), that is, how to quantify errors.

Q: What is true error?
A: True error denoted by E_t is the difference between the true value (also called the exact value) and the approximate value.

True Error = True value – Approximate value

Example 1

The derivative of a function $f(x)$ at a particular value of x can be approximately calculated by

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

of $f'(2)$ For $f(x) = 7e^{0.5x}$ and $h = 0.3$, find

a) the approximate value of $f'(2)$
b) the true value of $f'(2)$
c) the true error for part (a)

Solution

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

For $x = 2$ and $h = 0.3$,

$$f'(2) \approx \frac{f(2+0.3) - f(2)}{0.3}$$
$$= \frac{f(2.3) - f(2)}{0.3}$$
$$= \frac{7e^{0.5(2.3)} - 7e^{0.5(2)}}{0.3}$$
$$= \frac{22.107 - 19.028}{0.3}$$
$$= 10.265$$

b) The exact value of $f'(2)$ can be calculated by using our knowledge of differential calculus.

$$f(x) = 7e^{0.5x}$$
$$f'(x) = 7 \times 0.5 \times e^{0.5x}$$
$$= 3.5e^{0.5x}$$

So the true value of $f'(2)$ is

$$f'(2) = 3.5e^{0.5(2)}$$
$$= 9.5140$$

c) True error is calculated as

E_t = True value – Approximate value
$= 9.5140 - 10.265$
$= -0.75061$

The magnitude of true error does not show how bad the error is. A true error of $E_t = -0.722$ may seem to be small, but if the function given in the Example 1 were $f(x) = 7 \times 10^{-6} e^{0.5x}$, the true error in calculating $f'(2)$ with $h = 0.3$, would be $E_t = -0.75061 \times 10^{-6}$. This value of true error is smaller, even when the two problems are similar in that they use the same value of the function argument, $x = 2$ and the step size, $h = 0.3$. This brings us to the definition of relative true error.

Q: What is relative true error?
A: Relative true error is denoted by $\in_t$ and is defined as the ratio between the true error and the true value.

$$\text{Relative True Error} = \frac{\text{True Error}}{\text{True Value}}$$

Example 2

The derivative of a function $f(x)$ at a particular value of x can be approximately calculated by

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

For $f(x) = 7e^{0.5x}$ and $h = 0.3$, find the relative true error at $x = 2$.

Solution

From Example 1,

$$E_t = \text{True value} - \text{Approximate value}$$
$$= 9.5140 - 10.265$$
$$= -0.75061$$

Relative true error is calculated as

$$\in_t = \frac{\text{True Error}}{\text{True Value}}$$
$$= \frac{-0.75061}{9.5140}$$
$$= -0.078895$$

Relative true errors are also presented as percentages. For this example,

$$\in_t = -0.0758895 \times 100\%$$
$$= -7.58895\%$$

Absolute relative true errors may also need to be calculated. In such cases,

$$|\in_t| = |-0.075888|$$
$$= 0.0758895$$
$$= 7.58895\%$$

Q: What is approximate error?

A: In the previous section, we discussed how to calculate true errors. Such errors are calculated only if true values are known. An example where this would be useful is when one is checking if a program is in working order and you know some examples where the true error is known. But mostly we will not have the luxury of knowing true values as why would you want to find the approximate values if you know the true values. So when we are solving a problem numerically, we will only have access to approximate values. We need to know how to quantify error for such cases.

Approximate error is denoted by E_a and is defined as the difference between the present approximation and previous approximation.

Approximate Error = Present Approximation – Previous Approximation

Example 3

The derivative of a function $f(x)$ at a particular value of x can be approximately calculated by

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

For $f(x) = 7e^{0.5x}$ and at $x = 2$, find the following

a) $f'(2)$ using $h = 0.3$

b) $f'(2)$ using $h = 0.15$

c) approximate error for the value of $f'(2)$ for part (b)

Solution

a) The approximate expression for the derivative of a function is

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

For $x = 2$ and $h = 0.3$,

$$\begin{aligned} f'(2) &\approx \frac{f(2+0.3)-f(2)}{0.3} \\ &= \frac{f(2.3)-f(2)}{0.3} \\ &= \frac{7e^{0.5(2.3)}-7e^{0.5(2)}}{0.3} \\ &= \frac{22.107-19.028}{0.3} \\ &= 10.265 \end{aligned}$$

b) Repeat the procedure of part (a) with $h = 0.15$,

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

For $x = 2$ and $h = 0.15$,

$$\begin{aligned} f'(2) &\approx \frac{f(2+0.15)-f(2)}{0.15} \\ &= \frac{f(2.15)-f(2)}{0.15} \\ &= \frac{7e^{0.5(2.15)}-7e^{0.5(2)}}{0.15} \\ &= \frac{20.50-19.028}{0.15} \\ &= 9.8799 \end{aligned}$$

c) So the approximate error, E_a is

$$\begin{aligned} E_a &= \text{Present Approximation} - \text{Previous Approximation} \\ &= 9.8799 - 10.265 \\ &= -0.38474 \end{aligned}$$

The magnitude of approximate error does not show how bad the error is . An approximate error of $E_a = -0.38300$ may seem to be small; but for $f(x) = 7\times10^{-6}e^{0.5x}$, the approximate error in calculating $f'(2)$ with $h = 0.15$ would be $E_a = -0.38474\times10^{-6}$. This value of approximate error is smaller, even when the two problems are similar in that they use the same value of the function argument, $x = 2$, and $h = 0.15$ and $h = 0.3$. This brings us to the definition of relative approximate error.

Q: What is relative approximate error?

A: Relative approximate error is denoted by $\in_a$ and is defined as the ratio between the approximate error and the present approximation.

$$\text{Relative Approximate Error} = \frac{\text{Approximate Error}}{\text{Present Approximation}}$$

Example 4

The derivative of a function $f(x)$ at a particular value of x can be approximately calculated by

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

For $f(x) = 7e^{0.5x}$, find the relative approximate error in calculating $f'(2)$ using values from $h = 0.3$ and $h = 0.15$.

Solution

From Example 3, the approximate value of $f'(2) = 10.263$ using $h = 0.3$ and $f'(2) = 9.8800$ using $h = 0.15$.

$$\begin{aligned} E_a &= \text{Present Approximation} - \text{Previous Approximation} \\ &= 9.8799 - 10.265 \\ &= -0.38474 \end{aligned}$$

The relative approximate error is calculated as

$$\begin{aligned} \in_a &= \frac{\text{Approximate Error}}{\text{Present Approximation}} \\ &= \frac{-0.38474}{9.88799} \\ &= -0.038942 \end{aligned}$$

Relative approximate errors are also presented as percentages. For this example,

$$\begin{aligned} \in_a &= -0.038942 \times 100\% \\ &= -3.8942\% \end{aligned}$$

Absolute relative approximate errors may also need to be calculated. In this example

$$\begin{aligned} |\in_a| &= |-0.038942| \\ &= 0.038942 \text{ or } 3.8942\% \end{aligned}$$

Q: While solving a mathematical model using numerical methods, how can we use relative approximate errors to minimize the error?

A: In a numerical method that uses iterative methods, a user can calculate relative approximate error $\in_a$ at the end of each iteration. The user may pre-specify a minimum acceptable tolerance called the pre-specified tolerance, $\in_s$. If the absolute relative approximate error $\in_a$ is less than or equal to the pre-specified tolerance $\in_s$, that is, $|\in_a| \leq \in_s$, then the acceptable error has been reached and no more iterations would be required.

Alternatively, one may pre-specify how many significant digits they would like to be correct in their answer. In that case, if one wants at least m significant digits to be correct in the answer, then you would need to have the absolute relative approximate error, $|\in_a| \leq 0.5 \times 10^{2-m}$.

Q: But what do you mean by significant digits?

A: Significant digits are important in showing the truth one has in a reported number. For example, if someone asked me what the population of my county is, I would respond, "The

population of the Hillsborough county area is 1 million". But if someone was going to give me a $100 for every citizen of the county, I would have to get an exact count. That count would have been 1,079,587 in year 2003. So you can see that in my statement that the population is 1 million, that there is only one significant digit, that is, 1, and in the statement that the population is 1,079,587, there are seven significant digits. So, how do we differentiate the number of digits correct in 1,000,000 and 1,079,587? Well for that, one may use scientific notation. For our data we show

$$1{,}000{,}000 = 1 \times 10^6$$

$$1{,}079{,}587 = 1.079587 \times 10^6$$

to signify the correct number of significant digits.

Example 5

Give some examples of showing the number of significant digits.

Solution

a) 0.0459 has three significant digits
b) 4.590 has four significant digits
c) 4008 has four significant digits
d) 4008.0 has five significant digits
e) 1.079×10^3 has four significant digits
f) 1.0790×10^3 has five significant digits
g) 1.07900×10^3 has six significant digits

INTRODUCTION, APPROXIMATION AND ERRORS	
Topic	Measuring Errors
Summary	Textbook notes on measuring errors
Major	General Engineering
Authors	Autar Kaw
Date	December 7, 2008
Web Site	http://numericalmethods.eng.usf.edu

Multiple-Choice Test

Chapter 01.02
Measuring Errors

1. True error is defined as
   (A) Present Approximation – Previous Approximation
   (B) True Value – Approximate Value
   (C) abs (True Value – Approximate Value)
   (D) abs (Present Approximation – Previous Approximation)

2. The expression for true error in calculating the derivative of $\sin(2x)$ at $x = \pi/4$ by using the approximate expression
$$f'(x) \approx \frac{f(x+h) - f(x)}{h}$$
is
   (A) $\dfrac{h - \cos(2h) - 1}{h}$
   (B) $\dfrac{h - \cos(h) - 1}{h}$
   (C) $\dfrac{\cos(2h) - 1}{h}$
   (D) $\dfrac{\sin(2h)}{h}$

3. The relative approximate error at the end of an iteration to find the root of an equation is 0.004%. The least number of significant digits we can trust in the solution is
   (A) 2
   (B) 3
   (C) 4
   (D) 5

4. The number 0.01850×10^3 has _____ significant digits
   (A) 3
   (B) 4
   (C) 5
   (D) 6

5. The following gas stations were cited for irregular dispensation by the Department of Agriculture. Which one cheated you the most?

Station	Actual Gasoline dispensed	Gasoline Reading at pump
Ser	9.90	10.00
Cit	19.90	20.00
Hus	29.80	30.00
She	29.95	30.00

(A) Ser
(B) Cit
(C) Hus
(D) She

6. The number of significant digits in the number 219900 is
(A) 4
(B) 5
(C) 6
(D) 4 or 5 or 6

Answers

1. B
2. C
3. C
4. B
5. A
6. D

Problem Set

Chapter 01.02
Measuring Error

1. The trigonometric function $\sin(x)$ can be calculated by using the following infinite series

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + ...$$

a) What is the value of $\sin(2.17)$ by using the first three terms in the given series?
b) What is the value of $\sin(2.17)$ by using the first four terms in the given series?
c) Use your calculator for the true value of $\sin(2.17)$?
d) What is the true error for answer in part (a)?
e) What is the absolute true error for answer in part (a).
f) What is the relative true error for answer in part (a).
g) What is the absolute relative true error for answer in part (a).
h) What is the approximate error for answer in part (b)?
i) What is the absolute approximate error for answer in part (b).
j) What is the relative approximate error for answer in part (b).
k) What is the absolute relative approximate error for answer in part (b).
l) Assume that you do not know the exact value of $\sin(2.17)$, how many significant digits are at least correct if you use four terms in the series?
m) What should be the pre-specified relative error tolerance if at least 4 significant digits are required to be correct in calculating $\sin(2.17)$?

2. A Maclaurin series for a function is given by

$$f(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} + ...$$

What is the absolute relative approximate error if three terms are used for calculating $f(1.2)$?

3. A Maclaurin series for a function is given by

$$f(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} + ...$$

How many terms should be used in the series to consider that at least 2 significant digits are correct in your answer for $f(0.1)$?

4. A gas station owned by Valdez gives you 9.90 gallons of gasoline when you actually paid for 9.95 gallons. Another gas station owned by Hessup gives you 19.80 gallons of gasoline when you actually paid for 19.85 gallons of gasoline. If you only had these two gas stations available in your town, which one would you go to next time you had to fill up your car? Use the concepts learned in measuring errors to justify your answer.

5. The function e^x can be calculated by using the following infinite Maclaurin series

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + ...$$

   a) Use 5 terms to calculate the value of $e^{0.9}$?
   b) How many significant digits in my calculation would be correct if I use 5 terms?
   c) How do I know that I have used enough terms to calculate $e^{0.9}$, if I pre-specify a tolerance of 0.05%? What is the minimum number of terms I should use to achieve the pre-specified tolerance?
   d) Where are the sources of error coming from in the above series?

6. How many significant digits are correct in the following numbers
   a) 185000
   b) 0.0185
   c) 1.0185
   d) 185×10^3
   e) 1850×10^2
   f) 0.01850×10^5
   g) 0.0185×10^5
   h) 100.00
   i) 100.001

7. What is the correct normalized scientific notation for 0.029411765 with 4 significant digits?

Chapter 01.03
Sources of Error

After reading this chapter, you should be able to:

1. *know that there are two inherent sources of error in numerical methods – round-off and truncation error,*
2. *recognize the sources of round-off and truncation error, and*
3. *know the difference between round-off and truncation error.*

Error in solving an engineering or science problem can arise due to several factors. First, the error may be in the modeling technique. A mathematical model may be based on using assumptions that are not acceptable. For example, one may assume that the drag force on a car is proportional to the velocity of the car, but actually it is proportional to the square of the velocity of the car. This itself can create huge errors in determining the performance of the car, no matter how accurate the numerical methods you may use are. Second, errors may arise from mistakes in programs themselves or in the measurement of physical quantities. But, in applications of numerical methods itself, the two errors we need to focus on are

1. Round off error
2. Truncation error.

Q: What is round off error?

A: A computer can only represent a number approximately. For example, a number like $\frac{1}{3}$ may be represented as 0.333333 on a PC. Then the round off error in this case is $\frac{1}{3} - 0.333333 = 0.0000003\overline{3}$. Then there are other numbers that cannot be represented exactly. For example, π and $\sqrt{2}$ are numbers that need to be approximated in computer calculations.

Q: What problems can be created by round off errors?
A: Twenty-eight Americans were killed on February 25, 1991. An Iraqi Scud hit the Army barracks in Dhahran, Saudi Arabia. The patriot defense system had failed to track and intercept the Scud. What was the cause for this failure?
The Patriot defense system consists of an electronic detection device called the range gate. It calculates the area in the air space where it should look for a Scud. To find out where it

should aim next, it calculates the velocity of the Scud and the last time the radar detected the Scud. Time is saved in a register that has 24 bits length. Since the internal clock of the system is measured for every one-tenth of a second, 1/10 is expressed in a 24 bit-register as 0.00011001100110011001100. However, this is not an exact representation. In fact, it would need infinite numbers of bits to represent 1/10 exactly. So, the error in the representation in decimal format is

Figure 1 Patriot missile (Courtesy of the US Armed Forces, http://www.redstone.army.mil/history/archives/patriot/patriot.html)

$$\frac{1}{10}-(0\times 2^{-1}+0\times 2^{-2}+0\times 2^{-3}+1\times 2^{-4}+\ldots+1\times 2^{-22}+0\times 2^{-23}+0\times 2^{-24})$$
$$=9.537\times 10^{-8}$$

The battery was on for 100 consecutive hours, hence causing an inaccuracy of

$$=9.537\times 10^{-8}\frac{\text{s}}{0.1\text{s}}\times 100\text{ hr}\times\frac{3600\text{s}}{1\text{hr}}$$
$$=0.3433\text{s}$$

The shift calculated in the range gate due to 0.3433s was calculated as 687m. For the Patriot missile defense system, the target is considered out of range if the shift was going to more than 137m.

Q: What is truncation error?
A: Truncation error is defined as the error caused by truncating a mathematical procedure. For example, the Maclaurin series for e^x is given as

$$e^x=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\ldots\ldots\ldots\ldots$$

This series has an infinite number of terms but when using this series to calculate e^x, only a finite number of terms can be used. For example, if one uses three terms to calculate e^x, then

$$e^x\approx 1+x+\frac{x^2}{2!}.$$

the truncation error for such an approximation is

$$\text{Truncation error}=e^x-\left(1+x+\frac{x^2}{2!}\right),$$

$$= \frac{x^3}{3!} + \frac{x^4}{4!} + ...$$

But, how can truncation error be controlled in this example? We can use the concept of relative approximate error to see how many terms need to be considered. Assume that one is calculating $e^{1.2}$ using the Maclaurin series, then

$$e^{1.2} = 1 + 1.2 + \frac{1.2^2}{2!} + \frac{1.2^3}{3!} + ...$$

Let us assume one wants the absolute relative approximate error to be less than 1%. In Table 1, we show the value of $e^{1.2}$, approximate error and absolute relative approximate error as a function of the number of terms, n.

n	$e^{1.2}$	E_a	$\lvert\in_a\rvert\%$
1	1	-	-
2	2.2	1.2	54.546
3	2.92	0.72	24.658
4	3.208	0.288	8.9776
5	3.2944	0.0864	2.6226
6	3.3151	0.020736	0.62550

Using 6 terms of the series yields a $\lvert\in_a\rvert < 1\%$.

Q: Can you give me other examples of truncation error?

A: In many textbooks, the Maclaurin series is used as an example to illustrate truncation error. This may lead you to believe that truncation errors are just chopping a part of the series. However, truncation error can take place in other mathematical procedures as well. For example to find the derivative of a function, we define

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

But since we cannot use $\Delta x \to 0$, we have to use a finite value of Δx, to give

$$f'(x) \approx \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

So the truncation error is caused by choosing a finite value of Δx as opposed to a $\Delta x \to 0$.

For example, in finding $f'(3)$ for $f(x) = x^2$, we have the exact value calculated as follows.

$$f(x) = x^2$$
$$f'(x) = 2x$$
$$f'(3) = 2 \times 3$$
$$= 6$$

If we now choose $\Delta x = 0.2$, we get

$$f'(3) = \frac{f(3 + 0.2) - f(3)}{0.2}$$
$$= \frac{f(3.2) - f(3)}{0.2}$$

$$= \frac{3.2^2 - 3^2}{0.2}$$
$$= \frac{10.24 - 9}{0.2}$$
$$= \frac{1.24}{0.2}$$
$$= 6.2$$

We purposefully chose a simple function $f(x) = x^2$ with value of $x = 2$ and $\Delta x = 0.2$ because we wanted to have no round-off error in our calculations so that the truncation error can be isolated. The truncation error in this example is

$$6 - 6.2 = -0.2.$$

Can you reduce the truncate error by choosing a smaller Δx ?

Another example of truncation error is the numerical integration of a function,

$$I = \int_a^b f(x)dx$$

Exact calculations require us to calculate the area under the curve by adding the area of the rectangles as shown in Figure 2. However, exact calculations requires an infinite number of such rectangles. Since we cannot choose an infinite number of rectangles, we will have truncation error.

For example, to find

$$\int_3^9 x^2 dx,$$

we have the exact value as

$$\int_3^9 x^2 dx = \left[\frac{x^3}{3}\right]_3^9$$
$$= \left[\frac{9^3 - 3^3}{3}\right]$$
$$= 234$$

If we now choose to use two rectangles of equal width to approximate the area (see Figure 2) under the curve, the approximate value of the integral

$$\int_3^9 x^2 dx = (x^2)\Big|_{x=3}(6-3) + (x^2)\Big|_{x=6}(9-6)$$
$$= (3^2)3 + (6^2)3$$
$$= 27 + 108$$
$$= 135$$

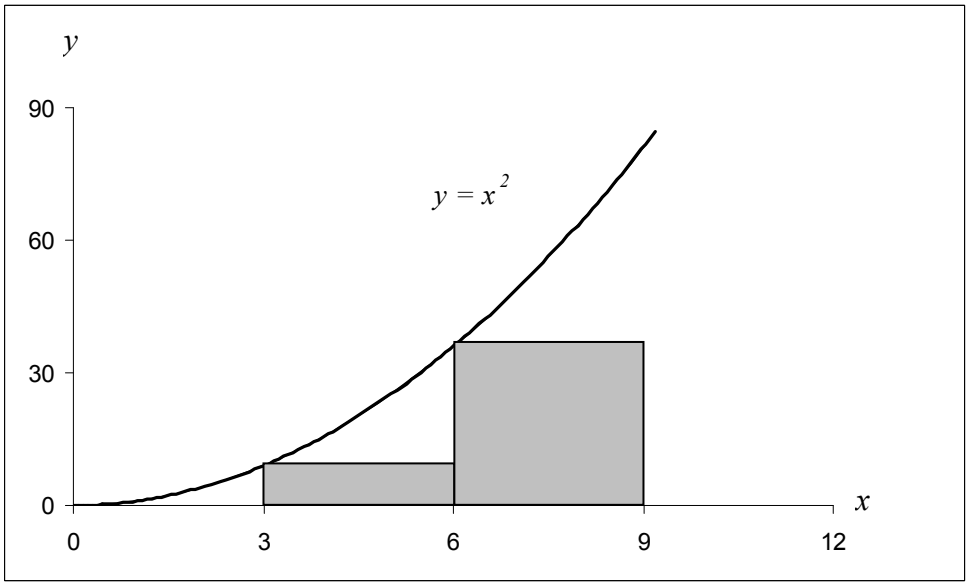


Figure 2 Plot of $y = x^2$ showing the approximate area under the curve from $x = 3$ to $x = 9$ using two rectangles.

Again, we purposefully chose a simple example because we wanted to have no round off error in our calculations. This makes the obtained error purely truncation. The truncation error is

$$234 - 135 = 99$$

Can you reduce the truncation error by choosing more rectangles as given in Figure 3? What is the truncation error?

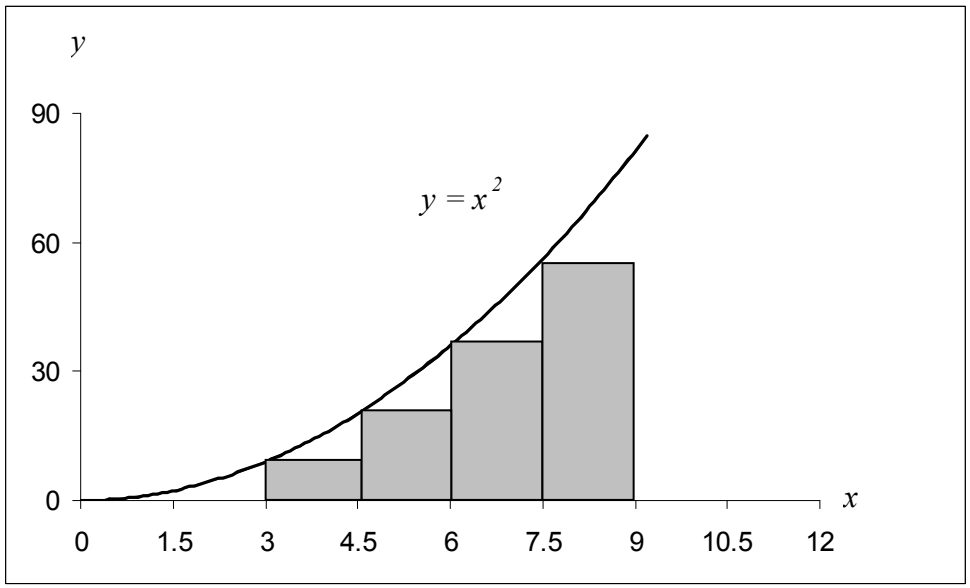


Figure 3 Plot of $y = x^2$ showing the approximate area under the curve from $x = 3$ to $x = 9$ using four rectangles.

References

"Patriot Missile Defense – Software Problem Led to System Failure at Dhahran, Saudi Arabia", GAO Report, General Accounting Office, Washington DC, February 4, 1992.

INTRODUCTION, APPROXIMATION AND ERRORS	
Topic	Sources of error
Summary	Textbook notes on sources of error
Major	General Engineering
Authors	Autar Kaw
Date	December 7, 2008
Web Site	http://numericalmethods.eng.usf.edu

Multiple-Choice Test

Chapter 01.03
Sources of Error

1. Truncation error is caused by approximating
   (A) irrational numbers
   (B) fractions
   (C) rational numbers
   (D) exact mathematical procedures.

2. A computer that represents only 4 significant digits with chopping would calculate 66.666*33.333 as
   (A) 2220
   (B) 2221
   (C) 2221.17778
   (D) 2222

3. A computer that represents only 4 significant digits with rounding off the last digit would calculate 66.666*33.333 as
   (A) 2220
   (B) 2221
   (C) 2221.17778
   (D) 2222

4. The truncation error in calculating $f'(2)$ for $f(x)=x^2$ by
$$f'(x) \cong \frac{f(x+h)-f(x)}{h}$$
with $h=0.2$ is
   (A) -0.2
   (B) 0.2
   (C) 4.0
   (D) 4.2

5. The truncation error in finding $\int_{-3}^{9} x^3 dx$ using LRAM (left end point Riemann approximation) with equally portioned points $-3<0<3<6<9$ is
   (A) 648
   (B) 810
   (C) 1620
   (D) 4200

6. The number 1/10 is registered in a fixed 6 bit-register with all bits used for the fractional part. The difference gets accumulated every $1/10^{th}$ of a second for one day. The magnitude of the accumulated difference in seconds is
   (A) 0.082
   (B) 135
   (C) 270
   (D) 540

Answers

1. D
2. B
3. D
4. B
5. A
6. D

Problem Set

Chapter 01.03
Sources of Error

1. What is the round off error in representing 200/3 in a 6-significant digit computer that chops the last significant digit?

2. What is the round off error in representing 200/3 in a 6-significant digit computer that rounds off the last significant digit?

3. What is the truncation error in the calculation of the $f'(x)$ that uses the approximation

$$f'(x) \approx \frac{f(x+\Delta x) - f(x)}{\Delta x}$$

for

$$f(x) = x^3, \quad \Delta x = 0.4, \text{ and } x = 5.$$

4. What is the truncation error in the calculation of $\sin(\pi/2)$ if only first five terms of the Maclaurin series are used for the calculation? Ignore the round off error in your calculations.

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + ...$$

5. The integral $\int_3^9 x^2 dx$ can be calculated approximately by finding the area of the four rectangles as shown in Figure 1. What is the truncation error due to this approximation?

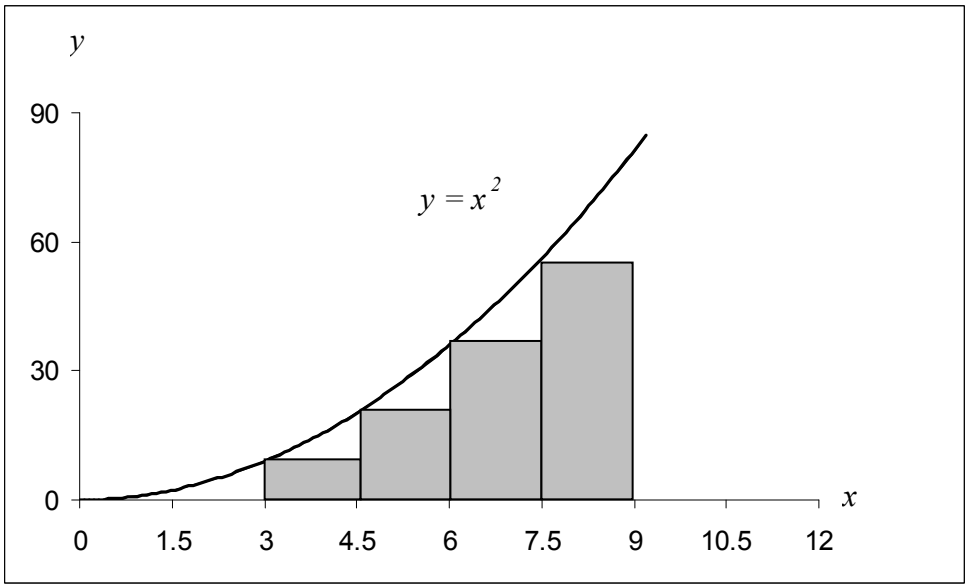


Figure 1 Plot of $y = x^2$ showing the approximate area under the curve from $x = 3$ to $x = 9$ using four rectangles

6. Below is the data given for thermal expansion coefficient of steel as a function of temperature.

Temperature	Instantaneous Thermal Expansion Coefficient
$^\circ$F	μ in/in$^\circ$F
80	6.47
60	6.36
40	6.24
20	6.12
0	6.00
-20	5.86
-40	5.72
-60	5.58
-80	5.43
-100	5.28
-120	5.09
-140	4.91
-160	4.72
-180	4.52
-200	4.30
-220	4.08
-240	3.83
-260	3.58
-280	3.33
-300	3.07

-320	2.76
-340	2.45

Regression is used to come up with a simple formula – a second order polynomial to relate the coefficient of thermal expansion coefficient of steel as a function of temperature. The formula given by MS Excel using default settings gives the formula as shown in Figure 2.

a) What is the value of the thermal expansion coefficient at $T = -300$, -200, -100, 0 and 100°F? Compare these values with those given in the table above.

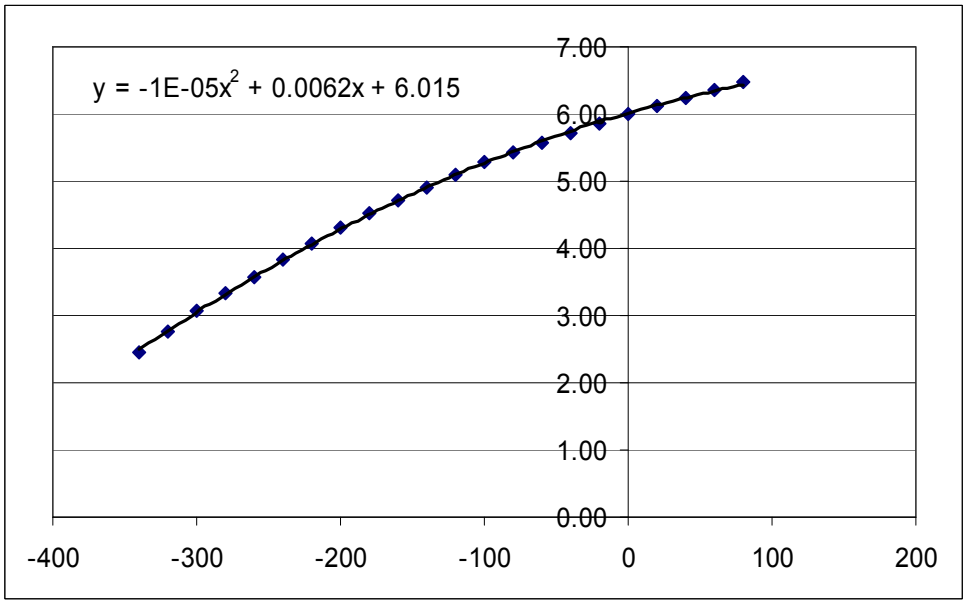


Figure 2 Default trend line given by Excel

b) Now the default settings for the trend line (Figure 3) are changed to scientific format with four significant digits. What is the value of the thermal expansion coefficient at $T = -300$, -200, -100, 0 and 100°F? Compare these values with those given in the table above. Do you get different answers in part (a) and part (b)? What do you attribute this difference to?

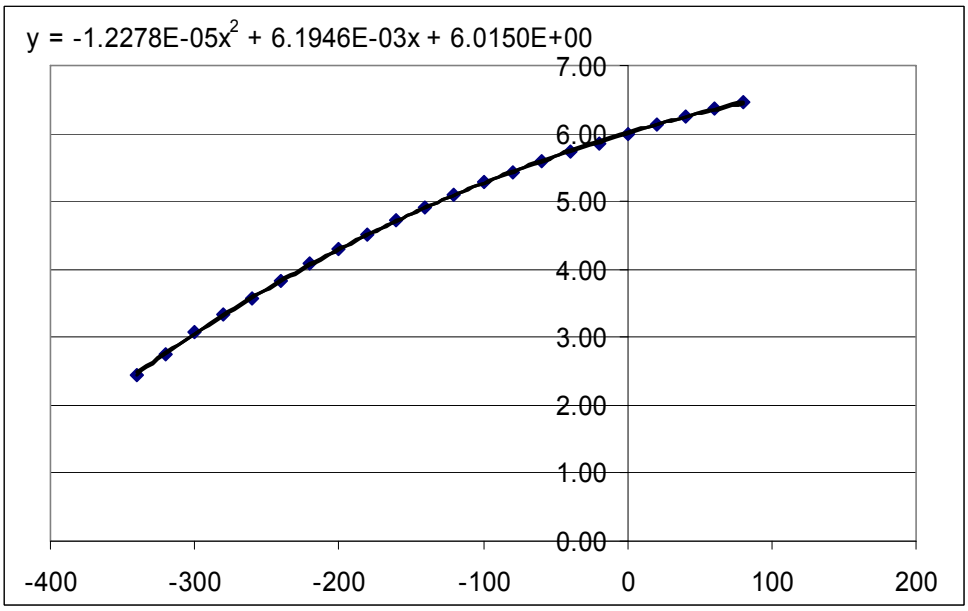


Figure 3 Trend line given by Excel using scientific format.

Chapter 01.04
Binary Representation of Numbers

After reading this chapter, you should be able to:

1. *convert a base-10 real number to its binary representation,*
2. *convert a binary number to an equivalent base-10 number.*

In everyday life, we use a number system with a base of 10. For example, look at the number 257.56. Each digit in 257.56 has a value of 0 through 9 and has a place value. It can be written as

$$257.76 = 2\times10^{2} + 5\times10^{1} + 7\times10^{0} + 7\times10^{-1} + 6\times10^{-2}$$

In a binary system, we have a similar system where the base is made of only two digits 0 and 1. So it is a base 2 system. A number like (1011.0011) in base-2 represents the decimal number as

$$\begin{aligned}(1011.0011)_2 &= \left((1\times2^{3} + 0\times2^{2} + 1\times2^{1} + 1\times2^{0}) + (0\times2^{-1} + 0\times2^{-2} + 1\times2^{-3} + 1\times2^{-4})\right)_{10} \\ &= 11.1875\end{aligned}$$

in the decimal system.

To understand the binary system, we need to be able to convert binary numbers to decimal numbers and vice-versa.

We have already seen an example of how binary numbers are converted to decimal numbers. Let us see how we can convert a decimal number to a binary number. For example take the decimal number 11.1875. First, look at the integer part: 11.

1. Divide 11 by 2. This gives a quotient of 5 and a remainder of 1. Since the remainder is 1, $a_0 = 1$.
2. Divide the quotient 5 by 2. This gives a quotient of 2 and a remainder of 1. Since the remainder is 1, $a_1 = 1$.
3. Divide the quotient 2 by 2. This gives a quotient of 1 and a remainder of 0. Since the remainder is 0, $a_2 = 0$.
4. Divide the quotient 1 by 2. This gives a quotient of 0 and a remainder of 1. Since the remainder is , $a_3 = 1$.

Since the quotient now is 0, the process is stopped. The above steps are summarized in Table 1.

Table 1 Converting a base-10 integer to binary representation.

	Quotient	Remainder
11/2	5	$1 = a_0$
5/2	2	$1 = a_1$
2/2	1	$0 = a_2$
1/2	0	$1 = a_3$

Hence

$$(11)_{10} = (a_3 a_2 a_1 a_0)_2$$
$$= (1011)_2$$

For any integer, the algorithm for finding the binary equivalent is given in the flow chart on the next page.

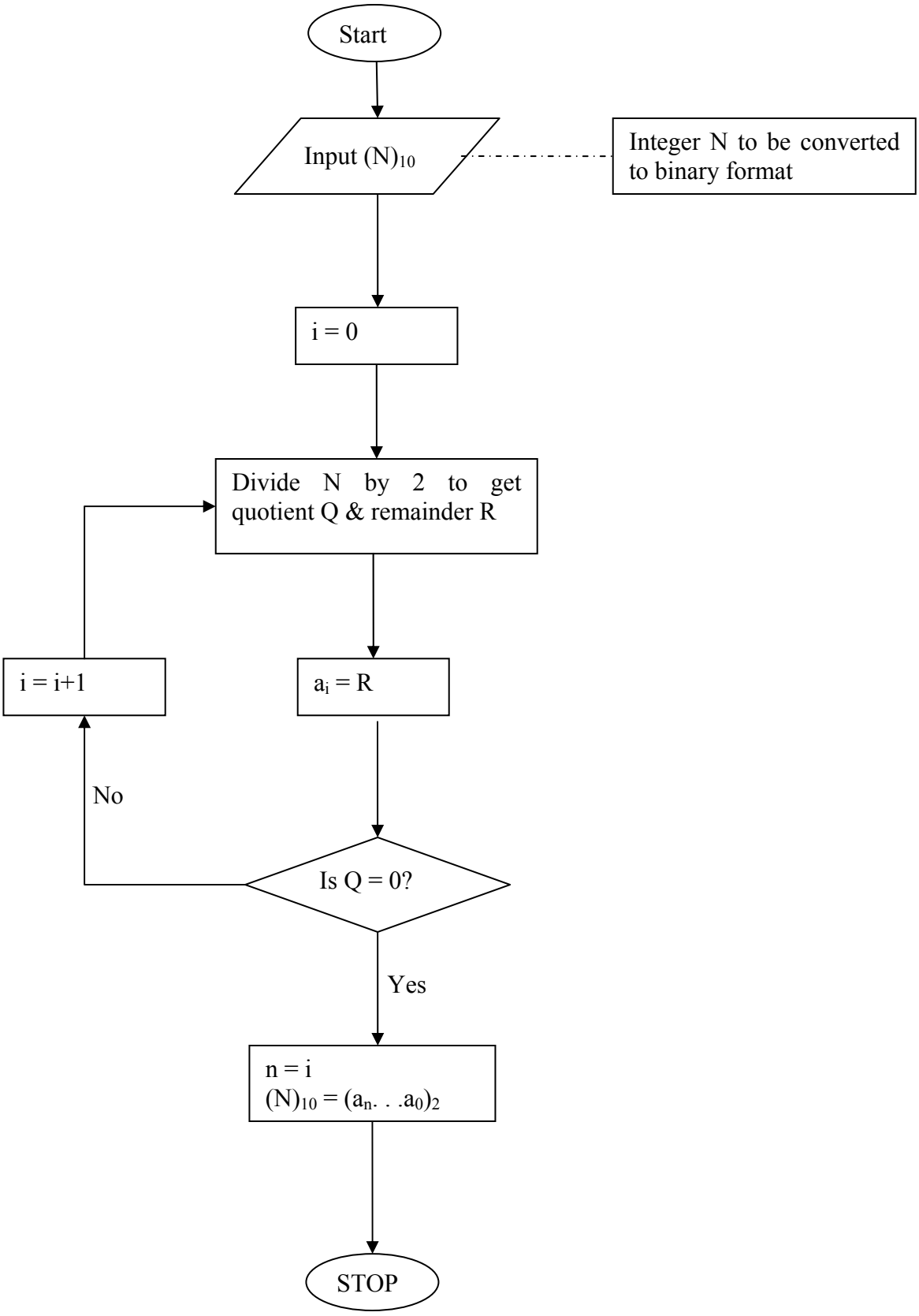
Now let us look at the decimal part, that is, 0.1875.

1. Multiply 0.1875 by 2. This gives 0.375. The number before the decimal is 0 and the number after the decimal is 0.375. Since the number before the decimal is 0, $a_{-1} = 0$.
2. Multiply the number after the decimal, that is, 0.375 by 2. This gives 0.75. The number before the decimal is 0 and the number after the decimal is 0.75. Since the number before the decimal is 0, $a_{-2} = 0$.
3. Multiply the number after the decimal, that is, 0.75 by 2. This gives 1.5. The number before the decimal is 1 and the number after the decimal is 0.5. Since the number before the decimal is 1, $a_{-3} = 1$.
4. Multiply the number after the decimal, that is, 0.5 by 2. This gives 1.0. The number before the decimal is 1 and the number after the decimal is 0. Since the number before the decimal is 1, $a_{-4} = 1$.

Since the number after the decimal is 0, the conversion is complete. The above steps are summarized in Table 2.

Table 2. Converting a base-10 fraction to binary representation.

	Number	Number after decimal	Number before decimal
0.1875×2	0.375	0.375	$0 = a_{-1}$
0.375×2	0.75	0.75	$0 = a_{-2}$
0.75×2	1.5	0.5	$1 = a_{-3}$
0.5×2	1.0	0.0	$1 = a_{-4}$


Start
Input $(N)_{10}$
Integer N to be converted to binary format
i = 0
Divide N by 2 to get quotient Q & remainder R
$a_i = R$
i = i+1
No
Is Q = 0?
Yes
n = i
$(N)_{10} = (a_n . . . a_0)_2$
STOP

Hence

$$(0.1875)_{10} = (a_{-1}a_{-2}a_{-3}a_{-4})_2$$
$$= (0.0011)_2$$

The algorithm for any fraction is given in a flowchart on the next page.
Having calculated

$$(11)_{10} = (1011)_2$$

and

$$(0.1875)_{10} = (0.0011)_2,$$

we have

$$(11.1875)_{10} = (1011.0011)_2.$$

In the above example, when we were converting the fractional part of the number, we were left with 0 after the decimal number and used that as a place to stop. In many cases, we are never left with a 0 after the decimal number. For example, finding the binary equivalent of 0.3 is summarized in Table 3.

Table 3. Converting a base-10 fraction to approximate binary representation.

	Number	Number after decimal	Number before decimal
0.3×2	0.6	0.6	$0 = a_{-1}$
0.6×2	1.2	0.2	$1 = a_{-2}$
0.2×2	0.4	0.4	$0 = a_{-3}$
0.4×2	0.8	0.8	$0 = a_{-4}$
0.8×2	1.6	0.6	$1 = a_{-5}$

As you can see the process will never end. In this case, the number can only be approximated in binary format, that is,

$$(0.3)_{10} \approx (a_{-1}a_{-2}a_{-3}a_{-4}a_{-5})_2 = (0.01001)_2$$

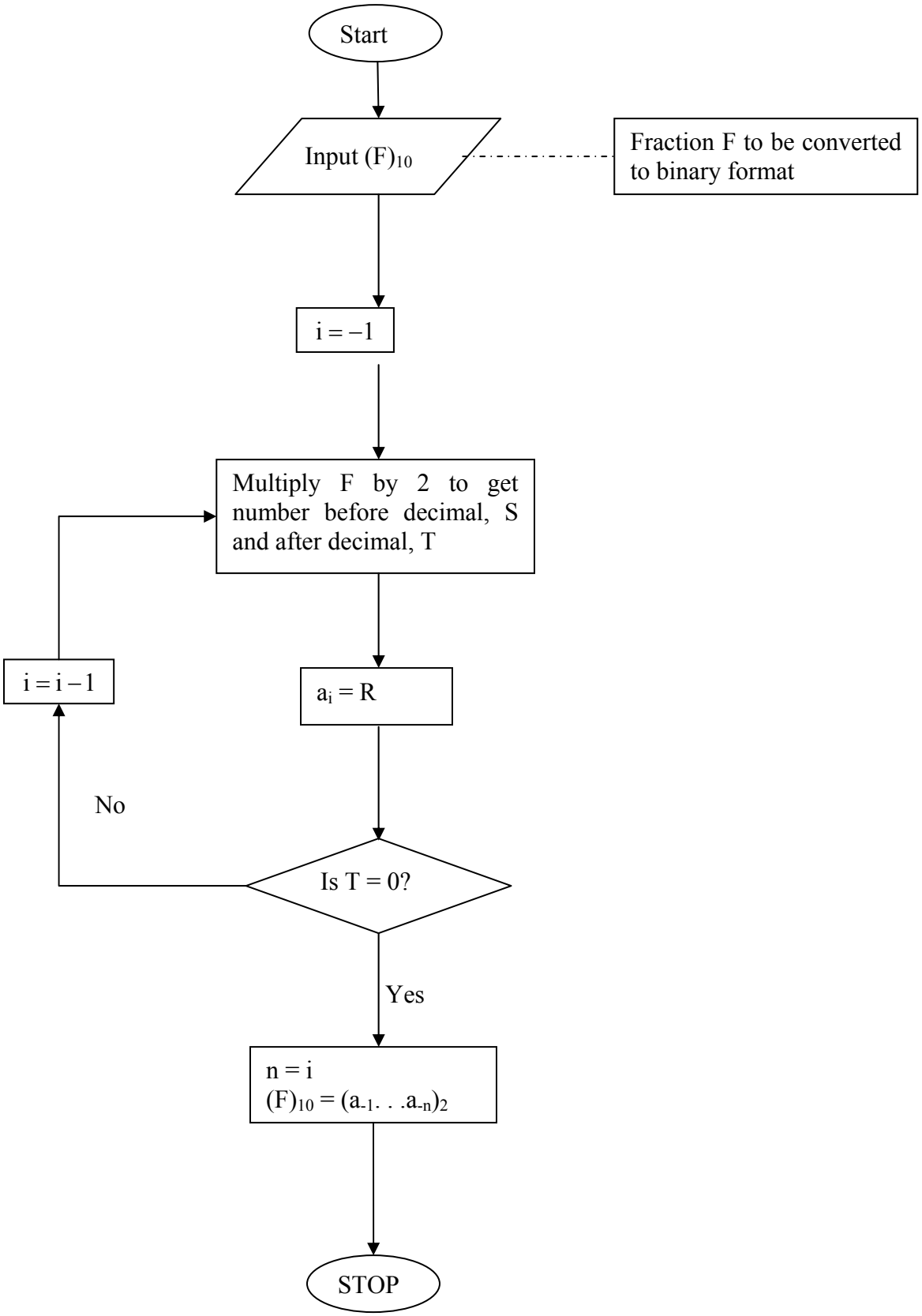
Q: But what is the mathematics behinds this process of converting a decimal number to binary format?
A: Let z be the decimal number written as

$$z = x.y$$

where

x is the integer part and y is the fractional part.

We want to find the binary equivalent of x. So we can write


Start
Input $(F)_{10}$
Fraction F to be converted to binary format
i = –1
Multiply F by 2 to get number before decimal, S and after decimal, T
$a_i = R$
i = i – 1
No
Is T = 0?
Yes
n = i
$(F)_{10} = (a_{-1}. . .a_{-n})_2$
STOP

$$x = a_n 2^n + a_{n-1} 2^{n-1} + ... + a_0 2^0$$

If we can now find $a_0, . . ., a_n$ in the above equation then

$$(x)_{10} = (a_n a_{n-1} . . . a_0)_2$$

We now want to find the binary equivalent of y. So we can write

$$y = b_{-1} 2^{-1} + b_{-2} 2^{-2} + ... + b_{-m} 2^{-m}$$

If we can now find $b_{-1}, . . ., b_{-m}$ in the above equation then

$$(y)_{10} = (b_{-1} b_{-2} . . . b_{-m})_2$$

Let us look at this using the same example as before.

Example 1

Convert $(11.1875)_{10}$ to base 2.

Solution

To convert $(11)_{10}$ to base 2, what is the highest power of 2 that is part of 11. That power is 3, as $2^3 = 8$ to give

$$11 = 2^3 + 3$$

What is the highest power of 2 that is part of 3. That power is 1, as $2^1 = 2$ to give

$$3 = 2^1 + 1$$

So

$$11 = 2^3 + 3 = 2^3 + 2^1 + 1$$

What is the highest power of 2 that is part of 1. That power is 0, as $2^0 = 1$ to give

$$1 = 2^0$$

Hence

$$(11)_{10} = 2^3 + 2^1 + 1 = 2^3 + 2^1 + 2^0 = 1 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 = (1011)_2$$

To convert $(0.1875)_{10}$ to the base 2, we proceed as follows. What is the smallest negative power of 2 that is less than or equal to 0.1875. That power is -3 as $2^{-3} = 0.125$.

So

$$0.1875 = 2^{-3} + 0.0625$$

What is the next smallest negative power of 2 that is less than or equal to 0.0625. That power is -4 as $2^{-4} = 0.0625$.

So

$$0.1875 = 2^{-3} + 2^{-4}$$

Hence

$$(0.1875)_{10} = 2^{-3} + 0.0625 = 2^{-3} + 2^{-4} = 0 \times 2^{-1} + 0 \times 2^{-2} + 1 \times 2^{-3} + 1 \times 2^{-4} = (0.0011)_2$$

Since

$$(11)_{10} = (1011)_2$$

and

$$(0.1875)_{10} = (0.0011)_2$$

we get

$$(11.1875)_{10} = (1011.0011)_2$$

Can you show this algebraically for any general number?

Example 2

Convert $(13.875)_{10}$ to base 2.

Solution

For $(13)_{10}$, conversion to binary format is shown in Table 4.

Table 4. Conversion of base-10 integer to binary format.

	Quotient	Remainder
13/2	6	$1 = a_0$
6/2	3	$0 = a_1$
3/2	1	$1 = a_2$
1/2	0	$1 = a_3$

So

$$(13)_{10} = (1101)_2 .$$

Conversion of $(0.875)_{10}$ to binary format is shown in Table 5.

Table 5. Converting a base-10 fraction to binary representation.

	Number	Number after decimal	Number before decimal
0.875×2	1.75	0.75	$1 = a_{-1}$
0.75×2	1.5	0.5	$1 = a_{-2}$
0.5×2	1.0	0.0	$1 = a_{-3}$

So

$$(0.875)_{10} = (0.111)_2$$

Hence

$$(13.875)_{10} = (1101.111)_2$$

INTRODUCTION TO NUMERICAL METHODS	
Topic	Binary representation of number
Summary	Textbook notes on binary representation of numbers
Major	General Engineering
Authors	Autar Kaw
Date	December 7, 2008
Web Site	http://numericalmethods.eng.usf.edu

Multiple-Choice Test

Chapter 01.04
Binary Representation

1. $(25)_{10} = (?)_2$
   (A) 100110
   (B) 10011
   (C) 11001
   (D) 110010

2. $(1101)_2 = (?)_{10}$
   (A) 3
   (B) 13
   (C) 15
   (D) 26

3. $(25.375)_{10} = (?.?)_2$
   (A) 100110.011
   (B) 11001.011
   (C) 10011.0011
   (D) 10011.110

4. Representing $\sqrt{2}$ in a fixed point register with 2 bits for the integer part and 3 bits for the fractional part gives a round off error of most nearly
   (A) -0.085709
   (B) 0.03921
   (C) 0.1642
   (D) 0.2892

5. An engineer working for the Department of Defense is writing a program that transfers non-negative real numbers to integer format. To avoid overflow problems, the maximum non-negative integer that can be represented in a 5-bit integer word is
   (A) 16
   (B) 31
   (C) 63
   (D) 64

6. For a numerically controlled machine, integers need to be stored in a memory location. The minimum number of bits needed for an integer word to represent all integers between 0 and 1024 is
   (A) 8
   (B) 9
   (C) 10
   (D) 11

Answers

1. C
2. B
3. B
4. B
5. B
6. D

Problem Set

Chapter 01.04
Binary Representation

1. Convert the following
   a) $(19)_{10} = (?)_2$
   b) $(75)_{10} = (?)_2$

2. Convert the following
   a) $(110111)_2 = (?)_{10}$
   b) $(11001)_2 = (?)_{10}$

3. Convert the following
   a) $(0.375)_{10} = (?)_2$
   b) $(0.075)_{10} = (?)_2$

4. Convert the following
   a) $(0.110001)_2 = (0.?)_{10}$
   b) $(0.0111)_2 = (0.?)_{10}$

5. Convert the following
   a) $(19.375)_{10} = (?.?)_2$
   b) $(75.075)_{10} = (?.?)_2$

6. Convert the following
   a) $(110111.110001)_2 = (?.?)_{10}$
   b) $(11001.0111)_2 = (?.?)_{10}$

Chapter 01.05
Floating Point Representation

After reading this chapter, you should be able to:

1. *convert a base-10 number to a binary floating point representation,*
2. *convert a binary floating point number to its equivalent base-10 number,*
3. *understand the IEEE-754 specifications of a floating point representation in a typical computer,*
4. *calculate the machine epsilon of a representation.*

Consider an old time cash register that would ring any purchase between 0 and 999.99 units of money. Note that there are five (not six) working spaces in the cash register (the decimal number is shown just for clarification).
Q: How will the smallest number 0 be represented?
A: The number 0 will be represented as

0	0	0	.	0	0

Q: How will the largest number 999.99 be represented?
A: The number 999.99 will be represented as

9	9	9	.	9	9

Q: Now look at any typical number between 0 and 999.99, such as 256.78. How would it be represented?
A: The number 256.78 will be represented as

2	5	6	.	7	8

Q: What is the smallest change between consecutive numbers?
A: It is 0.01, like between the numbers 256.78 and 256.79.

Q: What amount would one pay for an item, if it costs 256.789?
A: The amount one would pay would be rounded off to 256.79 or chopped to 256.78. In either case, the maximum error in the payment would be less than 0.01.

Q: What magnitude of relative errors would occur in a transaction?
A: Relative error for representing small numbers is going to be high, while for large numbers the relative error is going to be small.

For example, for 256.786, rounding it off to 256.79 accounts for a round-off error of $256.786 - 256.79 = -0.004$. The relative error in this case is

$$\varepsilon_t = \frac{-0.004}{256.786} \times 100$$
$$= -0.001558\%.$$

For another number, 3.546, rounding it off to 3.55 accounts for the same round-off error of $3.54 - 3.55 = -0.004$. The relative error in this case is

$$\varepsilon_t = \frac{-0.004}{3.546} \times 100$$
$$= -0.11280\%.$$

Q: If I am interested in keeping relative errors of similar magnitude for the range of numbers, what alternatives do I have?

A: To keep the relative error of similar order for all numbers, one may use a floating-point representation of the number. For example, in floating-point representation, a number

256.78 is written as $+2.5678 \times 10^2$,

0.003678 is written as $+3.678 \times 10^{-3}$, and

-256.789 is written as -2.56789×10^2.

The general representation of a number in base-10 format is given as

$$sign \times mantissa \times 10^{\text{exponent}}$$

or for a number y,

$$y = \sigma \times m \times 10^e$$

Where

σ = sign of the number, $+1$ or -1

m = mantissa, $1 \le m < 10$

e = integer exponent (also called ficand)

Let us go back to the example where we have five spaces available for a number. Let us also limit ourselves to positive numbers with positive exponents for this example. If we use the same five spaces, then let us use four for the mantissa and the last one for the exponent. So the smallest number that can be represented is 1 but the largest number would be 9.999×10^9. By using the floating-point representation, what we lose in accuracy, we gain in the range of numbers that can be represented. For our example, the maximum number represented changed from 999.99 to 9.999×10^9.

What is the error in representing numbers in the scientific format? Take the previous example of 256.78. It would be represented as 2.568×10^2 and in the five spaces as

2	5	6	8	2

Another example, the number 576329.78 would be represented as 5.763×10^5 and in five spaces as

5	7	6	3	5

So, how much error is caused by such representation. In representing 256.78, the round off error created is $256.78 - 256.8 = -0.02$, and the relative error is

$$\varepsilon_t = \frac{-0.02}{256.78} \times 100 = -0.0077888\%,$$

In representing 576329.78, the round off error created is $576329.78 - 5.763 \times 10^5 = 29.78$, and the relative error is

$$\varepsilon_t = \frac{29.78}{576329.78} \times 100 = 0.0051672\%.$$

What you are seeing now is that although the errors are large for large numbers, but the relative errors are of the same order for both large and small numbers.

Q: How does this floating-point format relate to binary format?
A: A number y would be written as

$$y = \sigma \times m \times 2^e$$

Where

σ = sign of number (negative or positive – use 0 for positive and 1 for negative),
m = mantissa, $(1)_2 \le m < (10)_2$, that is, $(1)_{10} \le m < (2)_{10}$, and
e = integer exponent.

Example 1

Represent $(54.75)_{10}$ in floating point binary format.

Solution

$$(54.75)_{10} = (110110.011)_2 = (1.10110011)_2 \times 2^{(5)_{10}}$$

The exponent 5 is equivalent in binary format as

$$(5)_{10} = (101)_2$$

Hence

$$(54.75)_{10} = (1.10110011)_2 \times 2^{(101)_2}$$

The sign of the number is positive, so

$$\sigma = 0$$

The mantissa

$m = 10110011$ (The leading 1 is not stored as it is always expected to be there),

and the exponent

$e = 101.$

Assuming that the number is written to a hypothetical word that is 9 bits long where the

- the first bit is used for the sign of the number,
- the second bit for the sign of the exponent,
- the next four bits for the mantissa (hence the mantissa is approximated as m=1011), and
- the next three bits for the exponent,

we have the representation as

0	0	1	0	1	1	1	0	1

Example 2

What number does the below given floating point format

0	1	1	0	1	1	1	1	0

represent in base-10 format. Assume a hypothetical 9-bit word, where the first bit is used for the sign of the number, second bit for the sign of the exponent, next four bits for the mantissa and next three for the exponent.

Solution

Given

Bit Representation	Part of Floating point number
0	Sign of number
1	Sign of exponent
1011	Magnitude of mantissa
110	Magnitude of exponent

The first bit is 0, so the number is positive.
The second bit is 1, so the exponent is negative.
The next four bits, 1011, are the magnitude of the mantissa, so

$$m = (1.1011)_2 = \left(1\times 2^0 + 1\times 2^{-1} + 0\times 2^{-2} + 1\times 2^{-3} + 1\times 2^{-4}\right)_{10} = (1.6875)_{10}$$

The last three bits, 110, are the magnitude of the exponent, so

$$e = (110)_2 = \left(1\times 2^2 + 1\times 2^1 + 0\times 2^0\right)_{10} = (6)_{10}$$

The number in binary format then is

$$(1.1011)_2 \times 2^{-(110)_2}$$

The number in base-10 format is

$$= 1.6875\times 2^{-6}$$
$$= 0.026367$$

Example 3

A machine stores floating-point numbers in a hypothetical 10-bit binary word. It employs the first bit for the sign of the number, the second one for the sign of the exponent, the next four for the exponent, and the last four for the magnitude of the mantissa.

a) Find how 0.02832 will be represented in the floating-point 10-bit word.
b) What is the decimal equivalent of the 10-bit word representation of part (a)?

Solution

a) For the number, we have the integer part as 0 and the fractional part as 0.02832
Let us first find the binary equivalent of the integer part

Integer part $(0)_{10} = (0)_2$

Now we find the binary equivalent of the fractional part

Fractional part:
$$.02832\times 2$$
$$0.05664\times 2$$
$$0.11328\times 2$$
$$0.22656\times 2$$
$$0.45312\times 2$$
$$0.90624\times 2$$

$$1.81248 \times 2$$
$$1.62496 \times 2$$
$$1.24992 \times 2$$
$$0.49984 \times 2$$
$$0.99968 \times 2$$
$$1.99936$$

Hence

$$(0.02832)_{10} \cong (0.00000111001)_2$$
$$= (1.11001)_2 \times 2^{-6}$$
$$\cong (1.1100)_2 \times 2^{-6}$$

The binary equivalent of exponent is found as follows

	Quotient	Remainder
6/2	3	$0 = a_0$
3/2	1	$1 = a_1$
1/2	0	$1 = a_2$

So

$$(6)_{10} = (110)_2$$

So

$$(0.02832)_{10} = (1.1100)_2 \times 2^{-(110)_2}$$
$$= (1.1100)_2 \times 2^{-(0110)_2}$$

Part of Floating point number	Bit Representation
Sign of number is positive	0
Sign of exponent is negative	1
Magnitude of the exponent	0110
Magnitude of mantissa	1100

The ten-bit representation bit by bit is

0	1	0	1	1	0	1	1	0	0

b) Converting the above floating point representation from part (a) to base 10 by following Example 2 gives

$$(1.1100)_2 \times 2^{-(0110)_2}$$
$$= \left(1 \times 2^0 + 1 \times 2^{-1} + 1 \times 2^{-2} + 0 \times 2^{-3} + 0 \times 2^{-4}\right) \times 2^{-\left(0 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 0 \times 2^0\right)}$$
$$= (1.75)_{10} \times 2^{-(6)_{10}}$$
$$= 0.02734375$$

Q: How do you determine the accuracy of a floating-point representation of a number?

A: The machine epsilon, $\in_{mach}$ is a measure of the accuracy of a floating point representation and is found by calculating the difference between 1 and the next number that can be represented. For example, assume a 10-bit hypothetical computer where the first bit is used

for the sign of the number, the second bit for the sign of the exponent, the next four bits for the exponent and the next four for the mantissa.
We represent 1 as

0	0	0	0	0	0	0	0	0	0

and the next higher number that can be represented is

0	0	0	0	0	0	0	0	0	1

The difference between the two numbers is

$$\begin{aligned}&(1.0001)_2 \times 2^{(0000)_2} - (1.0000)_2 \times 2^{(0000)_2} \\ &= (0.0001)_2 \\ &= (1 \times 2^{-4})_{10} \\ &= (0.0625)_{10}.\end{aligned}$$

The machine epsilon is

$$\in_{mach} = 0.0625.$$

The machine epsilon, $\in_{mach}$ is also simply calculated as two to the negative power of the number of bits used for mantissa. As far as determining accuracy, machine epsilon, $\in_{mach}$ is an upper bound of the magnitude of relative error that is created by the approximate representation of a number (See Example 4).

Example 4

A machine stores floating-point numbers in a hypothetical 10-bit binary word. It employs the first bit for the sign of the number, the second one for the sign of the exponent, the next four for the exponent, and the last four for the magnitude of the mantissa. Confirm that the magnitude of the relative true error that results from approximate representation of 0.02832 in the 10-bit format (as found in previous example) is less than the machine epsilon.

Solution

From Example 2, the ten-bit representation of 0.02832 bit-by-bit is

0	1	0	1	1	0	1	1	0	0

Again from Example 2, converting the above floating point representation to base-10 gives

$$\begin{aligned}&(1.1100)_2 \times 2^{-(0110)_2} \\ &= (1.75)_{10} \times 2^{-(6)_{10}} \\ &= (0.02734375)_{10}\end{aligned}$$

The absolute relative true error between the number 0.02832 and its approximate representation 0.02734375 is

$$\begin{aligned}|\varepsilon_t| &= \left|\frac{0.02832 - 0.02734375}{0.02832}\right| \\ &= 0.034472\end{aligned}$$

which is less than the machine epsilon for a computer that uses 4 bits for mantissa, that is,

$$\begin{aligned}\varepsilon_{mach} &= 2^{-4} \\ &= 0.0625.\end{aligned}$$

Q: How are numbers actually represented in floating point in a real computer?

A: In an actual typical computer, a real number is stored as per the IEEE-754 (Institute of Electrical and Electronics Engineers) floating-point arithmetic format. To keep the discussion short and simple, let us point out the salient features of the single precision format.

- A single precision number uses 32 bits.
- A number y is represented as

$$y = \sigma \times (1.a_1 a_2 \cdots a_{23}) \cdot 2^e$$

where

σ = sign of the number (positive or negative)

a_i = entries of the mantissa, can be only 0 or 1, $i = 1,..,23$

e =the exponent

Note the 1 before the radix point.

- The first bit represents the sign of the number (0 for positive number and 1 for a negative number).
- The next eight bits represent the exponent. Note that there is no separate bit for the sign of the exponent. The sign of the exponent is taken care of by normalizing by adding 127 to the actual exponent. For example in the previous example, the exponent was 6. It would be stored as the binary equivalent of $127 + 6 = 133$. Why is 127 and not some other number added to the actual exponent? Because in eight bits the largest integer that can be represented is $(11111111)_2 = 255$, and halfway of 255 is 127. This allows negative and positive exponents to be represented equally. The normalized (also called biased) exponent has the range from 0 to 255, and hence the exponent e has the range of $-127 \le e \le 128$.
- If instead of using the biased exponent, let us suppose we still used eight bits for the exponent but used one bit for the sign of the exponent and seven bits for the exponent magnitude. In seven bits, the largest integer that can be represented is $(1111111)_2 = 127$ in which case the exponent e range would have been smaller, that is, $-127 \le e \le 127$. By biasing the exponent, the unnecessary representation of a negative zero and positive zero exponent (which are the same) is also avoided.
- Actually, the biased exponent range used in the IEEE-754 format is not 0 to 255, but 1 to 254. Hence, exponent e has the range of $-126 \le e \le 127$. So what are $e = -127$ and $e = 128$ used for? If $e = 128$ and all the mantissa entries are zeros, the number is $\pm\infty$ (the sign of infinity is governed by the sign bit), if $e = 128$ and the mantissa entries are not zero, the number being represented is Not a Number (NaN). Because of the leading 1 in the floating point representation, the number zero cannot be represented exactly. That is why the number zero (0) is represented by $e = -127$ and all the mantissa entries being zero.
- The next twenty-three bits are used for the mantissa.
- The largest number by magnitude that is represented by this format is

$$\left(1 \times 2^0 + 1 \times 2^{-1} + 1 \times 2^{-2} + \cdots + 1 \times 2^{-22} + 1 \times 2^{-23}\right) \times 2^{127} = 3.40 \times 10^{38}$$

The smallest number by magnitude that is represented, other than zero, is

$$\left(1 \times 2^0 + 0 \times 2^{-1} + 0 \times 2^{-2} + \cdots + 0 \times 2^{-22} + 0 \times 2^{-23}\right) \times 2^{-126} = 1.18 \times 10^{-38}$$

- Since 23 bits are used for the mantissa, the machine epsilon,

$$\in_{mach} = 2^{-23}$$
$$= 1.19 \times 10^{-7}.$$

Q: How are numbers represented in floating point in double precision in a computer?
A: In double precision IEEE-754 format, a real number is stored in 64 bits.

- The first bit is used for the sign,
- the next 11 bits are used for the exponent, and
- the rest of the bits, that is 52, are used for mantissa.

Can you find in double precision the

- range of the biased exponent,
- smallest number that can be represented,
- largest number that can be represented, and
- machine epsilon?

INTRODUCTION TO NUMERICAL METHODS	
Topic	Floating Point Representation
Summary	Textbook notes on floating point representation
Major	General Engineering
Authors	Autar Kaw
Date	December 7, 2008
Web Site	http://numericalmethods.eng.usf.edu

Multiple-Choice Test

Chapter 01.05
Floating Point Representation

1. A hypothetical computer stores real numbers in floating point format in 8-bit words. The first bit is used for the sign of the number, the second bit for the sign of the exponent, the next two bits for the magnitude of the exponent, and the next four bits for the magnitude of the mantissa. The number $e \cong 2.718$ in the 8-bit format is
   (A) 00010101
   (B) 00011010
   (C) 00010011
   (D) 00101010

2. A hypothetical computer stores real numbers in floating point format in 8-bit words. The first bit is used for the sign of the number, the second bit for the sign of the exponent, the next two bits for the magnitude of the exponent, and the next four bits for the magnitude of the mantissa. The number 10100111 represented in the above given 8-bit format is
   (A) -5.75
   (B) -2.875
   (C) -1.75
   (D) -0.359375

3. A hypothetical computer stores floating point numbers in 8-bit words. The first bit is used for the sign of the number, the second bit for the sign of the exponent, the next two bits for the magnitude of the exponent, and the next four bits for the magnitude of the mantissa. The machine epsilon is most nearly
   (A) 2^{-7}
   (B) 2^{-4}
   (C) 2^{-3}
   (D) 2^{-2}

4. A machine stores floating point numbers in 7-bit word. The first bit is stored for the sign of the number, the next three for the biased exponent and the next three for the magnitude of the mantissa. The number (0010110) represented in base-10 is
   (A) 0.375
   (B) 0.875
   (C) 1.5
   (D) 3.5

5. A machine stores floating point numbers in 7-bit word. The first bit is stored for the sign of the number, the next three for the biased exponent and the next three for the magnitude of the mantissa. You are asked to represent 33.35 in the above word. The error you will get in this case would be
   (A) underflow
   (B) overflow
   (C) NaN
   (D) No error will be registered.

6. A hypothetical computer stores floating point numbers in 9-bit words. The first bit is used for the sign of the number, the second bit for the sign of the exponent, the next three bits for the magnitude of the exponent, and the next four bits for the magnitude of the mantissa. Every second, the error between 0.1 and its binary representation in the 9-bit word is accumulated. The accumulated error in seconds after one day most nearly is
   (A) 0.002344
   (B) 20.25
   (C) 202.5
   (D) 8640

Answers

1. A
2. A
3. B
4. B
5. B
6. C

Problem Set

Chapter 01.05
Floating Point Representation

1. A hypothetical computer stores real numbers in floating point format in 8-bit words. The first bit is used for the sign of the number, the second bit for the sign of the exponent, the next two bits for the magnitude of the exponent, and the next four bits for the magnitude of the mantissa. Represent 3.1415 in the 8-bit format.

2. A hypothetical computer stores real numbers in floating point format in 8-bit words. The first bit is used for the sign of the number, the second bit for the sign of the exponent, the next two bits for the magnitude of the exponent, and the next four bits for the magnitude of the mantissa. What number does 10101111 represent in the above given 8-bit format?

3. A hypothetical computer stores real numbers in floating point format in 10-bit words. The first bit is used for the sign of the number, the second bit for the sign of the exponent, the next three bits for the magnitude of the exponent, and the next five bits for the magnitude of the mantissa. Represent -0.0456 in the 10-bit format.

4. A hypothetical computer stores real numbers in floating point format in 10-bit words. The first bit is used for the sign of the number, the second bit for the sign of the exponent, the next three bits for the magnitude of the exponent, and the next five bits for the magnitude of the mantissa. What number does 1011010011 represent in the above given 10-bit format?

5. A machine stores floating point numbers in 7-bit words. Employ first bit for the sign of the number, second one for the sign of the exponent, next two for the magnitude of the exponent, and the last three for the magnitude of the mantissa.
    a) By magnitude, what are the smallest negative and positive numbers in the system?
    b) By magnitude, what are the largest negative and positive numbers in the system?
    c) What is the machine epsilon?
    d) Represent e^1 in the 7-bit format.
    e) Represent 3.623 in the 7-bit format.
    f) What is the next higher number, x_2 after $x_1 = 0\ 1\ 1\ 0\ 1\ 1\ 0$ in the 7-bit format.
    g) Find $\left|\frac{x_2 - x_1}{x_1}\right|$ from part (f) and compare with the machine epsilon.

Chapter 01.06
Propagation of Errors

If a calculation is made with numbers that are not exact, then the calculation itself will have an error. How do the errors in each individual number propagate through the calculations. Let's look at the concept via some examples.

Example 1

Find the bounds for the propagation error in adding two numbers. For example if one is calculating $X + Y$ where

$X = 1.5 \pm 0.05$,

$Y = 3.4 \pm 0.04$.

Solution

By looking at the numbers, the maximum possible value of X and Y are

$X = 1.55$ and $Y = 3.44$

Hence

$$X + Y = 1.55 + 3.44 = 4.99$$

is the maximum value of $X + Y$.

The minimum possible value of X and Y are

$X = 1.45$ and $Y = 3.36$.

Hence

$$X + Y = 1.45 + 3.36$$
$$= 4.81$$

is the minimum value of $X + Y$.

Hence

$$4.81 \le X + Y \le 4.99.$$

One can find similar intervals of the bound for the other arithmetic operations of $X - Y, X * Y,$ and X / Y . What if the evaluations we are making are function evaluations instead? How do we find the value of the propagation error in such cases.

If f is a function of several variables $X_1, X_2, X_3, ..., X_{n-1}, X_n$, then the maximum possible value of the error in f is

$$\Delta f \approx \left|\frac{\partial f}{\partial X_1}\Delta X_1\right| + \left|\frac{\partial f}{\partial X_2}\Delta X_2\right| + ... + \left|\frac{\partial f}{\partial X_{n-1}}\Delta X_{n-1}\right| + \left|\frac{\partial f}{\partial X_n}\Delta X_n\right|$$

Example 2

The strain in an axial member of a square cross-section is given by

$$\in = \frac{F}{h^2 E}$$

where

F =axial force in the member, N
h = length or width of the cross-section, m
E =Young's modulus, Pa

Given

$F = 72 \pm 0.9$ N
$h = 4 \pm 0.1$ mm
$E = 70 \pm 1.5$ GPa

Find the maximum possible error in the measured strain.

Solution

$$\in = \frac{72}{(4\times10^{-3})^2(70\times10^9)}$$
$$= 64.286\times10^{-6}$$
$$= 64.286\mu$$
$$\Delta \in = \left|\frac{\partial \in}{\partial F}\Delta F\right| + \left|\frac{\partial \in}{\partial h}\Delta h\right| + \left|\frac{\partial \in}{\partial E}\Delta E\right|$$
$$\frac{\partial \in}{\partial F} = \frac{1}{h^2 E}$$
$$\frac{\partial \in}{\partial h} = -\frac{2F}{h^3 E}$$
$$\frac{\partial \in}{\partial E} = -\frac{F}{h^2 E^2}$$
$$\Delta E = \left|\frac{1}{h^2 E}\Delta F\right| + \left|\frac{2F}{h^3 E}\Delta h\right| + \left|\frac{F}{h^2 E^2}\Delta E\right|$$
$$= \left|\frac{1}{(4\times10^{-3})^2(70\times10^9)}\times 0.9\right| + \left|\frac{2\times 72}{(4\times10^{-3})^3(70\times10^9)}\times 0.0001\right| + \left|\frac{72}{(4\times10^{-3})^2(70\times10^9)^2}\times 1.5\times10^9\right|$$
$$= 8.0357\times10^{-7} + 3.2143\times10^{-6} + 1.3776\times10^{-6}$$
$$= 5.3955\times10^{-6}$$
$$= 5.3955\mu$$

Hence

$$\in = (64.286\mu \pm 5.3955\mu)$$

implying that the axial strain, $\in$ is between 58.8905μ and 69.6815μ

Example 3

Subtraction of numbers that are nearly equal can create unwanted inaccuracies. Using the formula for error propagation, show that this is true.

Solution

Let

$$z = x - y$$

Then

$$|\Delta z| = \left|\frac{\partial z}{\partial x}\Delta x\right| + \left|\frac{\partial z}{\partial y}\Delta y\right|$$
$$= |(1)\Delta x| + |(-1)\Delta y|$$
$$= |\Delta x| + |\Delta y|$$

So the absolute relative change is

$$\left|\frac{\Delta z}{z}\right| = \frac{|\Delta x| + |\Delta y|}{|x - y|}$$

As x and y become close to each other, the denominator becomes small and hence create large relative errors.

For example if

$$x = 2 \pm 0.001$$
$$y = 2.003 \pm 0.001$$
$$\left|\frac{\Delta z}{z}\right| = \frac{|0.001| + |0.001|}{|2 - 2.003|}$$
$$= 0.6667$$
$$= 66.67\%$$

INTRODUCTION TO NUMERICAL METHODS	
Topic	Propagation of Errors
Summary	Textbook notes on how errors propagate in arithmetic and function evaluations
Major	All Majors of Engineering
Authors	Autar Kaw
Last Revised	December 7, 2008
Web Site	http://numericalmethods.eng.usf.edu

Multiple-Choice Test

Chapter 01.06
Propagation of Errors

1. If $A = 3.56 \pm 0.05$ and $B = 3.25 \pm 0.04$, the values of $A + B$ are
   (A) $6.81 \le A + B \le 6.90$
   (B) $6.72 \le A + B \le 6.90$
   (C) $6.81 \le A + B \le 6.81$
   (D) $6.71 \le A + B \le 6.91$

2. A number A is correctly rounded to 3.18 from a given number B. Then $|A - B| \le C$, where C is
   (A) 0.005
   (B) 0.01
   (C) 0.18
   (D) 0.09999

3. Two numbers A and B are approximated as C and D, respectively. The relative error in $C \times D$ is given by
   (A) $\left|\left(\frac{A-C}{A}\right)\right| \times \left|\left(\frac{B-D}{B}\right)\right|$
   (B) $\left|\left(\frac{A-C}{A}\right)\right| + \left|\left(\frac{B-D}{B}\right)\right| + \left|\left(\frac{A-C}{A}\right)\right| \times \left|\left(\frac{B-D}{B}\right)\right|$
   (C) $\left|\left(\frac{A-C}{A}\right)\right| + \left|\left(\frac{B-D}{B}\right)\right| - \left|\left(\frac{A-C}{A}\right)\right| \times \left|\left(\frac{B-D}{B}\right)\right|$
   (D) $\left(\frac{A-C}{A}\right) - \left(\frac{B-D}{B}\right)$

4. The formula for normal strain in a longitudinal bar is given by $\in = \frac{F}{AE}$ where
   F = normal force applied
   A = cross-sectional area of the bar
   E = Young's modulus
   If $F = 50 \pm 0.5\text{N}$, $A = 0.2 \pm 0.002\ \text{m}^2$, and $E = 210 \times 10^9 \pm 1 \times 10^9\ \text{Pa}$, the maximum error in the measurement of strain is
   (A) 10^{-12}
   (B) 2.95×10^{-11}
   (C) 1.22×10^{-9}
   (D) 1.19×10^{-9}

5. A wooden block is measured to be 60 mm by a ruler and the measurements are considered to be good to 1/4th of a millimeter. Then in the measurement 60 mm, we have _____ significant digits
   (A) 0
   (B) 1
   (C) 2
   (D) 3

6. In the calculation of the volume of a cube of nominal size 5", the uncertainty in the measurement of each side is 10%. The uncertainty in the measurement of the volume would be
   (A) 5.477%
   (B) 10.00%
   (C) 17.32%
   (D) 30.00%

Answers

1. B
2. A
3. C
4. B
5. C
6. D

Chapter 01.07
Taylor Theorem Revisited

After reading this chapter, you should be able to

1. *understand the basics of Taylor's theorem,*
2. *write transcendental and trigonometric functions as Taylor's polynomial,*
3. *use Taylor's theorem to find the values of a function at any point, given the values of the function and all its derivatives at a particular point,*
4. *calculate errors and error bounds of approximating a function by Taylor series, and*
5. *revisit the chapter whenever Taylor's theorem is used to derive or explain numerical methods for various mathematical procedures.*

The use of Taylor series exists in so many aspects of numerical methods that it is imperative to devote a separate chapter to its review and applications. For example, you must have come across expressions such as

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots \tag{1}$$

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \tag{2}$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \tag{3}$$

All the above expressions are actually a special case of Taylor series called the Maclaurin series. Why are these applications of Taylor's theorem important for numerical methods? Expressions such as given in Equations (1), (2) and (3) give you a way to find the approximate values of these functions by using the basic arithmetic operations of addition, subtraction, division, and multiplication.

Example 1

Find the value of $e^{0.25}$ using the first five terms of the Maclaurin series.

Solution

The first five terms of the Maclaurin series for e^x is

$$e^x \approx 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!}$$

$$e^{0.25} \approx 1+0.25+\frac{0.25^2}{2!}+\frac{0.25^3}{3!}+\frac{0.25^4}{4!}$$
$$=1.2840$$

The exact value of $e^{0.25}$ up to 5 significant digits is also 1.2840.
But the above discussion and example do not answer our question of what a Taylor series is. Here it is, for a function $f(x)$

$$f(x+h)=f(x)+f'(x)h+\frac{f''(x)}{2!}h^2+\frac{f'''(x)}{3!}h^3+\cdots \quad (4)$$

provided all derivatives of $f(x)$ exist and are continuous between x and $x+h$.

What does this mean in plain English?

As Archimedes would have said (*without the fine print*), "*Give me the value of the function at a single point, and the value of all (first, second, and so on) its derivatives, and I can give you the value of the function at any other point*".

It is very important to note that the Taylor series is not asking for the expression of the function and its derivatives, just the value of the function and its derivatives at a single point.

Now the fine print: Yes, all the derivatives have to exist and be continuous between x (the point where you are) to the point, $x+h$ where you are wanting to calculate the function at. However, if you want to calculate the function approximately by using the n^{th} order Taylor polynomial, then $1^{st}, 2^{nd}, ..., n^{th}$ derivatives need to exist and be continuous in the closed interval $[x, x+h]$, while the $(n+1)^{th}$ derivative needs to exist and be continuous in the open interval $(x, x+h)$.

Example 2

Take $f(x)=\sin(x)$, we all know the value of $\sin\left(\frac{\pi}{2}\right)=1$. We also know the $f'(x)=\cos(x)$ and $\cos\left(\frac{\pi}{2}\right)=0$. Similarly $f''(x)=-\sin(x)$ and $\sin\left(\frac{\pi}{2}\right)=1$. In a way, we know the value of $\sin(x)$ and all its derivatives at $x=\frac{\pi}{2}$. We do not need to use any calculators, just plain differential calculus and trigonometry would do. Can you use Taylor series and this information to find the value of $\sin(2)$?

Solution

$$x=\frac{\pi}{2}$$
$$x+h=2$$
$$h=2-x$$
$$=2-\frac{\pi}{2}$$
$$=0.42920$$

So

$$f(x+h)=f(x)+f'(x)h+f''(x)\frac{h^2}{2!}+f'''(x)\frac{h^3}{3!}+f''''(x)\frac{h^4}{4!}+\cdots$$

$$x=\frac{\pi}{2}$$

$$h=0.42920$$

$$f(x)=\sin(x),\ f\left(\frac{\pi}{2}\right)=\sin\left(\frac{\pi}{2}\right)=1$$

$$f'(x)=\cos(x),\ f'\left(\frac{\pi}{2}\right)=0$$

$$f''(x)=-\sin(x),\ f'''\left(\frac{\pi}{2}\right)=-1$$

$$f'''(x)=-\cos(x),\ f'''\left(\frac{\pi}{2}\right)=0$$

$$f''''(x)=\sin(x),\ f''''\left(\frac{\pi}{2}\right)=1$$

Hence

$$f\left(\frac{\pi}{2}+h\right)=f\left(\frac{\pi}{2}\right)+f'\left(\frac{\pi}{2}\right)h+f''\left(\frac{\pi}{2}\right)\frac{h^2}{2!}+f'''\left(\frac{\pi}{2}\right)\frac{h^3}{3!}+f''''\left(\frac{\pi}{2}\right)\frac{h^4}{4!}+\cdots$$

$$f\left(\frac{\pi}{2}+0.42920\right)=1+0(0.42920)-1\frac{(0.42920)^2}{2!}+0\frac{(0.42920)^3}{3!}+1\frac{(0.42920)^4}{4!}+\cdots$$

$$=1+0-0.092106+0+0.00141393+\cdots$$

$$\cong 0.90931$$

The value of $\sin(2)$ I get from my calculator is 0.90930 which is very close to the value I just obtained. Now you can get a better value by using more terms of the series. In addition, you can now use the value calculated for $\sin(2)$ coupled with the value of $\cos(2)$ (which can be calculated by Taylor series just like this example or by using the $\sin^2 x+\cos^2 x\equiv 1$ identity) to find value of $\sin(x)$ at some other point. In this way, we can find the value of $\sin(x)$ for any value from $x=0$ to 2π and then can use the periodicity of $\sin(x)$, that is $\sin(x)=\sin(x+2n\pi), n=1,2,\ldots$ to calculate the value of $\sin(x)$ at any other point.

Example 3

Derive the Maclaurin series of $\sin(x)=x-\frac{x^3}{3!}+\frac{x^5}{5!}-\frac{x^7}{7!}+\cdots$

Solution

In the previous example, we wrote the Taylor series for $\sin(x)$ around the point $x=\frac{\pi}{2}$. Maclaurin series is simply a Taylor series for the point $x=0$.

$$f(x)=\sin(x),\ f(0)=0$$

$$f'(x)=\cos(x), f'(0)=1$$
$$f''(x)=-\sin(x), f''(0)=0$$
$$f'''(x)=-\cos(x), f'''(0)=-1$$
$$f''''(x)=\sin(x), f''''(0)=0$$
$$f'''''(x)=\cos(x), f'''''(0)=1$$

Using the Taylor series now,

$$f(x+h)=f(x)+f'(x)h+f''(x)\frac{h^2}{2!}+f'''(x)\frac{h^3}{3!}+f''''(x)\frac{h^4}{4}+f'''''(x)\frac{h^5}{5}+\cdots$$
$$f(0+h)=f(0)+f'(0)h+f''(0)\frac{h^2}{2!}+f'''(0)\frac{h^3}{3!}+f''''(0)\frac{h^4}{4}+f'''''(0)\frac{h^5}{5}+\cdots$$
$$f(h)=f(0)+f'(0)h+f''(0)\frac{h^2}{2!}+f'''(0)\frac{h^3}{3!}+f''''(0)\frac{h^4}{4}+f'''''(0)\frac{h^5}{5}+\cdots$$
$$=0+1(h)-0\frac{h^2}{2!}-1\frac{h^3}{3!}+0\frac{h^4}{4}+1\frac{h^5}{5}+\cdots$$
$$=h-\frac{h^3}{3!}+\frac{h^5}{5!}+\cdots$$

So

$$f(x)=x-\frac{x^3}{3!}+\frac{x^5}{5!}-\cdots$$
$$\sin(x)=x-\frac{x^3}{3!}+\frac{x^5}{5!}-\cdots$$

Example 4

Find the value of $f(6)$ given that $f(4)=125$, $f'(4)=74$, $f''(4)=30$, $f'''(4)=6$ and all other higher derivatives of $f(x)$ at $x=4$ are zero.

Solution

$$f(x+h)=f(x)+f'(x)h+f''(x)\frac{h^2}{2!}+f'''(x)\frac{h^3}{3!}+\cdots$$
$$x=4$$
$$h=6-4$$
$$=2$$

Since fourth and higher derivatives of $f(x)$ are zero at $x=4$.

$$f(4+2)=f(4)+f'(4)2+f''(4)\frac{2^2}{2!}+f'''(4)\frac{2^3}{3!}$$
$$f(6)=125+74(2)+30\left(\frac{2^2}{2!}\right)+6\left(\frac{2^3}{3!}\right)$$
$$=125+148+60+8$$
$$=341$$

Note that to find $f(6)$ exactly, we only needed the value of the function and all its derivatives at some other point, in this case, $x=4$. We did not need the expression for the function and all its derivatives. Taylor series application would be redundant if we needed to know the expression for the function, as we could just substitute $x=6$ in it to get the value of $f(6)$.

Actually the problem posed above was obtained from a known function $f(x)=x^3+3x^2+2x+5$ where $f(4)=125$, $f'(4)=74$, $f''(4)=30$, $f'''(4)=6$, and all other higher derivatives are zero.

Error in Taylor Series

As you have noticed, the Taylor series has infinite terms. Only in special cases such as a finite polynomial does it have a finite number of terms. So whenever you are using a Taylor series to calculate the value of a function, it is being calculated approximately.

The Taylor polynomial of order n of a function $f(x)$ with $(n+1)$ continuous derivatives in the domain $[x, x+h]$ is given by

$$f(x+h)=f(x)+f'(x)h+f''(x)\frac{h^2}{2!}+\cdots+f^{(n)}(x)\frac{h^n}{n!}+R_n(x)$$

where the remainder is given by

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

where

$$x<c<x+h$$

that is, c is some point in the domain $(x, x+h)$.

Example 5

The Taylor series for e^x at point $x=0$ is given by

$$e^x=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\frac{x^4}{4!}+\frac{x^5}{5!}+\cdots$$

a) What is the truncation (true) error in the representation of e^1 if only four terms of the series are used?
b) Use the remainder theorem to find the bounds of the truncation error.
Solution

a) If only four terms of the series are used, then

$$e^x\approx 1+x+\frac{x^2}{2!}+\frac{x^3}{3!}$$

$$e^1\approx 1+1+\frac{1^2}{2!}+\frac{1^3}{3!}$$

$$=2.66667$$

The truncation (true) error would be the unused terms of the Taylor series, which then are

$$E_t = \frac{x^4}{4!} + \frac{x^5}{5!} + \cdots$$
$$= \frac{1^4}{4!} + \frac{1^5}{5!} + \cdots$$
$$\cong 0.0516152$$

b) But is there any way to know the bounds of this error other than calculating it directly? Yes,

$$f(x+h) = f(x) + f'(x)h + \cdots + f^{(n)}(x)\frac{h^n}{n!} + R_n(x)$$

where

$$R_n(x) = \frac{(x-h)^{n+1}}{(n+1)!} f^{(n+1)}(c),\ x < c < x+h, \text{ and}$$

c is some point in the domain $(x, x+h)$. So in this case, if we are using four terms of the Taylor series, the remainder is given by $(x=0, n=3)$

$$R_3(x) = \frac{(0-1)^{3+1}}{(3+1)!} f^{(3+1)}(c)$$
$$= \frac{1}{4!} f^{(4)}(c)$$
$$= \frac{e^c}{24}$$

Since

$$x < c < x+h$$
$$0 < c < 0+1$$
$$0 < c < 1$$

The error is bound between

$$\frac{e^0}{24} < R_3(1) < \frac{e^1}{24}$$
$$\frac{1}{24} < R_3(1) < \frac{e}{24}$$
$$0.041667 < R_3(1) < 0.113261$$

So the bound of the error is less than 0.113261 which does concur with the calculated error of 0.0516152.

Example 6

The Taylor series for e^x at point $x=0$ is given by

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \cdots$$

As you can see in the previous example that by taking more terms, the error bounds decrease and hence you have a better estimate of e^1. How many terms it would require to get an approximation of e^1 within a magnitude of true error of less than 10^{-6}?

Solution

Using $(n+1)$ terms of the Taylor series gives an error bound of

$$R_n(x) = \frac{(x-h)^{n+1}}{(n+1)!} f^{(n+1)}(c)$$

$$x = 0, h = 1, f(x) = e^x$$

$$R_n(0) = \frac{(0-1)^{n+1}}{(n+1)!} f^{(n+1)}(c)$$

$$= \frac{(-1)^{n+1}}{(n+1)!} e^c$$

Since

$$x < c < x + h$$

$$0 < c < 0 + 1$$

$$0 < c < 1$$

$$\frac{1}{(n+1)!} < |R_n(0)| < \frac{e}{(n+1)!}$$

So if we want to find out how many terms it would require to get an approximation of e^1 within a magnitude of true error of less than 10^{-6},

$$\frac{e}{(n+1)!} < 10^{-6}$$

$$(n+1)! > 10^6 e$$

$$(n+1)! > 10^6 \times 3 \qquad \text{(as we do not know the value of } e \text{ but it is less than 3).}$$

$$n \geq 9$$

So 9 terms or more will get e^1 within an error of 10^{-6} in its value.

We can do calculations such as the ones given above only for simple functions. To do a similar analysis of how many terms of the series are needed for a specified accuracy for any general function, we can do that based on the concept of absolute relative approximate errors discussed in Chapter 01.02 as follows.

We use the concept of absolute relative approximate error (see Chapter 01.02 for details), which is calculated after each term in the series is added. The maximum value of m, for which the absolute relative approximate error is less than $0.5 \times 10^{2-m}\%$ is the least number of significant digits correct in the answer. It establishes the accuracy of the approximate value of a function without the knowledge of remainder of Taylor series or the true error.

INTRODUCTION TO NUMERICAL METHODS	
Topic	Taylor Theorem Revisited
Summary	These are textbook notes on Taylor Series
Major	All engineering majors
Authors	Autar Kaw
Date	December 7, 2008
Web Site	http://numericalmethods.eng.usf.edu

Multiple-Choice Test

Chapter 01.07
Taylors Series Revisited

1. The coefficient of the x^5 term in the Maclaurin polynomial for $\sin(2x)$ is
   (A) 0
   (B) 0.0083333
   (C) 0.016667
   (D) 0.26667

2. Given $f(3)=6$, $f'(3)=8$, $f''(3)=11$, and that all other higher order derivatives of $f(x)$ are zero at $x=3$, and assuming the function and all its derivatives exist and are continuous between $x=3$ and $x=7$, the value of $f(7)$ is
   (A) 38.000
   (B) 79.500
   (C) 126.00
   (D) 331.50

3. Given that $y(x)$ is the solution to $\frac{dy}{dx}=y^3+2$, $y(0)=3$ the value of $y(0.2)$ from a second order Taylor polynomial is
   (A) 4.400
   (B) 8.800
   (C) 24.46
   (D) 29.00

4. The series $\sum_{n=0}^{\infty}(-1)\frac{x^{2n}}{(2n)!}4^n$ is a Maclaurin series for the following function
   (A) $\cos(x)$
   (B) $\cos(2x)$
   (C) $\sin(x)$
   (D) $\sin(2x)$

5. The function

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

is called the error function. It is used in the field of probability and cannot be calculated exactly for finite values of x. However, one can expand the integrand as a Taylor polynomial and conduct integration. The approximate value of $erf(2.0)$ using first three terms of the Taylor series around $t = 0$ is

(A) -0.75225
(B) 0.99532
(C) 1.5330
(D) 2.8586

6. Using the remainder of Maclaurin polynomial of n^{th} order for $f(x)$ defined as

$$R_n(x) = \frac{x^{n+1}}{(n+1)!} f^{(n+1)}(c), \quad n \geq 0, \ 0 \leq c \leq x$$

the least order of the Maclaurin polynomial required to get an absolute true error of at most 10^{-6} in the calculation of $\sin(0.1)$ is (do not use the exact value of $\sin(0.1)$ or $\cos(0.1)$ to find the answer, but the knowledge that $|\sin(x)| \leq 1$ and $|\cos(x)| \leq 1$).

(A) 3
(B) 5
(C) 7
(D) 9

Answers

1. D
2. C
3. C
4. B
5. A
6. B