COMPUTATION
AND
PROBLEM SOLVING
IN
UNDERGRADUATE PHYSICS
Second Edition

IDL       MATLAB       OCTAVE
PYTHON    MAXIMA       MAPLE
• MATHEMATICA FORTRAN C
NUMERICAL RECIPES LSODE MUDPACK
• Б\TeX      TGF

DAVID M. COOK

Department of Physics
Lawrence University
711 E Boldt Way, SPC 24
Appleton, Wisconsin 54911

Copyright © First Edition 2000–16 by David M. Cook
Copyright © Second Edition 2016–23 by David M. Cook

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (creativecommons.org/licenses/by-nc-sa/4.0/). Any use not permitted by this license requires authorization in writing from David M. Cook.
The first edition of this publication was registered with the Library of Congress with the call number QC20.C66.2004.

The second edition was registered with the Library of Congress with the LCCN 2019947884. It has been assigned the ISBN 9780961342968 and carries the Library of Congress call number QC20.C66.2019.
Preface to the Second Edition

Note: Usually with a second edition of a book, the preface to the first edition\textsuperscript{1} is preserved and a short preface to the second edition is added. In the present case, the preface to the first edition required here and there a number of edits which, had they not been made, would have been perhaps a bit confusing to readers of the second edition. Consequently, I have elected to depart from the normal practice and simply create a preface for the second edition, though much of its content essentially copies that of the preface to the first edition.

Note: Regardless of which components are included and which omitted in this version of Computation and Problem Solving in Undergraduate Physics, the preface, acknowledgements, and disclaimer in the front matter are those from the assemblage containing all components.

Since the mid 1980’s (including the years since my official retirement in 2008), we in the Department of Physics at Lawrence University have been developing and offering curricular components that

- support efforts to acquaint students with computational procedures and resources early enough so that they will be motivated and prepared to use these resources on their own initiative when circumstances warrant and so that later work need not be interrupted to deal with computational issues as an aside to its main purposes, and

- provide students with both the background and the confidence to support informed reading of vendor manuals, which usually do a splendid job of listing capabilities exhaustively but typically burden the beginner with initially irrelevant refinements and fail to illustrate adequately how even the rudimentary capabilities can be combined to perform useful tasks.

Over the years since the mid 1980s, a wide assortment of instructional materials has been drafted and redrafted. This book brings these materials together into a single publication with the hope that it may prove useful to others who seek to achieve these same or similar objectives.\textsuperscript{2}

With these objectives in mind, this book consciously focuses on helping students get started. It is not designed to be comprehensive or exhaustive, either in laying out the capabilities of any particular computational resource or in discussing numerical algorithms. Students must understand throughout that they must refer regularly to vendors’ manuals and on-line help files for details beyond those discussed in the book—details that may, in fact, be necessary for successful completion of some of the exercises. The need for that activity is noted here; repeated reminders will not be included in the body of the book.

The book is also not a book about computational physics; it addresses uses of computational tools. Indeed, the sophomore course at Lawrence in which students first encounter this book would not in any way replace a course in computational physics. Rather, the materials treated here should

\textsuperscript{1}The first edition was published in 2003, though it experienced a number of edits and adjustments in subsequent years.

\textsuperscript{2}The article by David M. Cook titled “Computation in Undergraduate Physics: The Lawrence Approach” and appearing in the American Journal of Physics (Am. J. Phys. 76, 321–326 (April-May 2008) describes these efforts in some detail.
provide strong background for a subsequent, junior-senior level course in computational physics, which would—I believe—be substantially enhanced if students came to it already familiar with the resources on which this book concentrates.

One major difficulty in creating materials on computational topics is that different potential users favor different hardware platforms and software packages. Especially in the computational arena, the variety of options and combinations is so great that any single choice (or coordinated set of choices) is bound to limit the usefulness of the product to a small subset of all potential users. This book addresses that difficulty by being assembled from a wide assortment of components, some of which—the generic components—will be included in all versions and others of which—those specific to particular software packages—will be included only if the potential user requests them. Thus, the specific software and hardware discussed in the book can be tailored to the spectrum of resources available at the instructor’s site. Two versions may well differ in numerous respects. One may include the generic components and the components that discuss\(^3\) IDL, MAPLE, C (with Numerical Recipes), and $\LaTeX$ while another may include the generic components and the components that focus on MATLAB, Mathematica, and FORTRAN (including Numerical Recipes). The table of contents and index contain only entries from the included chapters and sections. To facilitate communication among users of different versions, however, chapter and section numbers and the numbers identifying package-independent exercises are preserved in all versions. In a version that does not include FORTRAN, for example, the FORTRAN sections will be omitted from Chapters 9, 10, 11, 13, 14, and 15 and Chapters 12 and 16—for which FORTRAN is prerequisite—will be omitted altogether. In addition, FORTRAN-specific exercises will be omitted from the end-of-chapter exercises. Because of version-specific omissions such as those just described, there will therefore be gaps in the chapter, section, and exercise numbers in any version that does not include all options. In contrast, within each chapter, equation numbers, figure numbers, table numbers, and footnote numbers advance from one without gaps, and page numbers run continuously from the beginning of the book to the end. In consequence, the numbers assigned to identical equations, figures, tables, footnotes, and pages may differ from version to version, but the numbers assigned to chapters and sections with identical titles and the numbers assigned to identical exercises will be the same in all versions (and will have gaps reflecting omitted chapters, sections, and exercises). Such flexibility would be impossible were we not able to exploit features of $\LaTeX$, including the particular capabilities of the \texttt{ifthen} and \texttt{imakeidx} packages, to assemble the PDF files that permit the selected assemblage subsequently to be viewed on a screen or printed.

Even among sites that use the same spectrum of hardware and software, however, some aspects of local environments remain unique to individual sites. Local rules of citizenship; the features and elementary resources of the local operating system; local practices and policies governing structuring of public directories, assignment of accounts and passwords, backup schedules, and after-hours access; licensing restrictions on proprietary software; means to launch particular application programs, compile user-written FORTRAN and/or C programs, and access printers; and numerous other aspects are subject to considerable local variation. This book does not constrain local options in these matters. Instead, its users must draft a site-specific supplement, which we will refer to as the Local Guide, to which individuals should refer for site-specific particulars. A $\LaTeX$ template for that guide, specifically the one used at Lawrence, is available to users of this book, but it will require editing to reflect local practices. In particular, to give local sites flexibility in configuring their environments, we have in the book used symbols like $\$HEAD$, $\$IDLHEAD$, and $\$NRHEAD$ to stand for paths to the specific directories that sit at the head of particular directory trees. All such symbols must be expanded as described in the Local Guide when commands or statements illustrated in the book are submitted to the user’s machine.

With the broadest brush, Chapter 1 stands alone and focuses on a number of topics assumed as background for the rest of the book. The next several chapters introduce\(^4\)

\(^3\)Many of the packages mentioned in this list are commercial and proprietary, and the names are registered trademarks of the respective vendors. Full contact information for all mentioned packages will be found in Appendix Z.

\(^4\) The second edition has added the shareware programs OCTAVE and PYTHON and replaced the no-longer available commercial program MACSYMA with the shareware program MAXIMA. The addition of OCTAVE and
• Specific array processors (Chapter 2 on IDL, Chapter 3 on MATLAB, Chapter 4 on OCTAVE, Chapter 5 on PYTHON),
• Computer algebra systems (Chapter 6 on MAXIMA, Chapter 7 on MAPLE, Chapter 8 on Mathematica),
• Programming languages (Chapter 9—with sections on FORTRAN and C), and
• Subroutine libraries (Chapter 10 on Numerical Recipes, Chapter 12 on LSODE, Chapter 16 on MUDPACK).

The remaining chapters address several important categories of computational processing, specifically

• Solving ordinary differential equations (Chapters 11 and 12),
• Evaluating integrals (Chapter 13),
• Finding roots (Chapter 14),
• Solving partial differential equations (Chapters 15 and 16)

Each of Chapters 11, 13, 14, and 15 begins with a (generic) section in which several problems drawn from subareas of physics and using the computational technique on which the chapter focuses are laid out. Each of Chapters 11, 13, and 14 then continues with

• one or more (optional) sections in which some of the identified problems are addressed with whatever computer algebra systems are included in the version,
• a (generic) section on numerical approaches to the category of problem on which the chapter focuses, and
• several (optional) sections in which some of the problems laid out in the first section are addressed with whatever array processors, computer algebra systems, and programming languages are included in the version.

Somewhat in contrast, Chapter 15 continues with

• a (generic) section on finite difference methods (FDMs) for solving partial differential equations,
• several (optional) sections in which some of the identified problems are addressed using FDMs with each of several tools,
• a (generic) section on finite element methods (FEMs), and
• several (optional) sections in which some of the identified problems are addressed using FEMs.

Chapter 12 begins with a brief orientation to LSODE—a large and well-tested publicly available FORTRAN program for solving systems of ordinary differential equations—and Chapter 16 begins with a general discussion of multigrid techniques for solving partial differential equations and then provides an orientation to MUDPACK—a large collection of publicly available FORTRAN programs for solving partial differential equations using those techniques. Each chapter then illustrates the use of the tools on which it focuses to address several representative problems. Every chapter in the book concludes with a collection of exercises using the techniques—both symbolic and numerical—of the chapter. The appendices introduce a publishing system (Appendix A on \LaTeX) and a (UNIX/LINUX) program for producing drawings (Appendix B on TGIF).

The order of presentation in the book does not compel any particular order of treatment in a course or program of self-study. To be sure, some later sections depend on some earlier sections, but the linkages are not particularly tight. In the Lawrence context, for example, the required sophomore course Computational Mechanics typically covers the chapters and appendices introducing either IDL or MATLAB, either MAPLE or Mathematica, and \LaTeX; and finally covers either the IDL or MATLAB and either the MAPLE or the Mathematica portions of the chapters on ordinary PYTHON explain why Chapters 4–12 in the first edition have become Chapters 6–14 in the second edition.
differential equations (ODEs), integration and root finding. The chapter on Numerical Recipes, the
FORTRAN and/or C portions of the chapters on programming, ODEs, integration, and root finding,
the chapter on partial differential equations and the chapters on LSODE and MUDPACK are the
focus of the Lawrence elective junior/senior course Computational Physics.

Despite the organization of the chapters by program or by computational technique involved,
the focus throughout is on physical contexts. The materials are designed to be used in conjunction
with intermediate level courses, not introductory courses. While the illustrations of computational
procedures highlight significant physical contexts and most of the examples and suggested exercises
emerge from interesting physical situations, the objective is for students to become both fluent
and wary in using computational resources in application to these physical situations, not to dwell
excessively on the microscopic details of numerical analysis or to teach them the underlying physics
(except insofar as successful computer-based solution of problems underscores the power of the
fundamental physical ideas). The students are assumed

- to have completed an introductory survey course in physics,
- to have completed courses in calculus, differential equations and, to some extent, linear algebra,
  and
- to be embarking on intermediate-level studies in physics

as they undertake a study of this book. We focus not so much on the set up of the situations—that
is assumed to be the province of other courses—as on computer-based techniques and strategies for
determining the solution once the set up is complete. Examples are drawn from classical mechanics,
classical electricity and magnetism, thermal physics, quantum mechanics, curve fitting, DC and AC
circuit theory, optics, and several other areas.

David M. Cook
Appleton, Wisconsin
14 February 2023
Acknowledgements

First and foremost, I wish to acknowledge the assistance and contributions of Peter Strunk, LU ’89, Kristi R. G. Hendrickson, LU ’91, Todd G. Ruskell, LU ’91, Stephen L. Mielke, LU ’92, Ruth Rhodes, LU ’92, Michelle Ruprecht, LU ’92, Sandra Collins, LU ’93, Mark F. Gehreke, LU ’93, Karl J. Geissler, LU ’94, Steven Van Metre, LU ’94, Alain Bellon, LU ’95, Peter Kelly Senecal, LU ’95, Christopher C. Schmidt, LU ’97, Michael D. Stenner, LU ’97, Mark Nornberg, LU ’98, Scot Shaw, LU ’98, Jim Truitt, LU ’98, Eric D. Moore, LU ’99, Teresa K. Hayne, LU ’00, Danica Dralus, LU ’02, Ryan T. Peterson, LU ’03, Scott J. Kaminski, LU ’04, Michelle L. Milne, LU ’04, Lauren E. Kost, LU ’05, Claire Weiss, LU ’07, and Erik Garbackik, LU ’08—all Lawrence students who, as undergraduates, contributed during the first decade and a half to the evolution of this book and the accompanying manual of solutions to representative exercises. While current versions in most instances deviate—often considerably—from the initial drafts to which these students contributed, there can be no denying that their contributions have played an important role in the evolution of the use of computational resources in the Lawrence curriculum and the evolution of this book. I thank them warmly and sincerely for their assistance. All of these students have given permission for me to use whatever has over the years become of their contributions in this publication.

Second, I wish to acknowledge and thank several individuals who have provided reviews of drafts or otherwise assisted in the refinement of this book, including

- numerous students (beyond those specifically named in the previous paragraph) who have enrolled in my courses over the years and who, directly and indirectly, have made comments that have influenced this book.
- my departmental colleagues J. Bruce Brackenridge, John R. Brandenberger, Jeffrey A. Collett, Matthew R. Stoneking, and Paul W. Fontana, who have listened to my sometimes lengthy observations about matters computational and curricular and offered valuable criticisms.
- Wolfgang Christian (Davidson College), Robert Ehrlich (George Mason University), and A. John Mallinckrodt (California Polytechnic Institute–Pomona), each of whom participated in a 1990 Sloan-supported conference on computation in advanced undergraduate physics that I organized at Lawrence and each of whom as a consultant throughout the writing project has offered many suggestions and constructive criticisms.
- all those faculty members\(^1\) who participated in one of the NSF-supported week-long workshops offered at Lawrence in the summers of 2001, 2002, and 2003 and who, in that context,

---

\(^1\)In July 2001: Steve Adams (Widener University), Russel Kauffman (Muhlenberg College), Marylin Bell (Lake Forest College), Lawrence A. Molnar (Calvin College), Thomas R. Greenlee (Bethel College), Elliott Moore (New Mexico Tech), Javier Hasbun (State University of West Georgia), Joelle L. Murray (Linfield College), Derrick Hylton (Spelman College), Paula C. Turner (Kenyon College), Ross Hyman (DePaul University), Scott N. Walck (Lebanon Valley College), William H. Ingham (James Madison University), and Tim Young (University of North Dakota). In July, 2002 (first workshop): Albert Batten (United States Air Force Academy), Andrea Cox LU ’91 (Beloit College), Brian Cudnik (Prairie View A and M), Craig Gunsul (Whitman College), Kevin Lee (University of Nebraska), Kuan-Chi Luk (University of California, Berkeley), Mary Macomber (Boise State University), Viktor Martisovits (Central College), Donald Miller (Central Missouri State University), Dorn Peterson (James Madison University), Robert Ragan (University of Wisconsin – Lacrosse), Shafiq Rahman (Allegheny College), Rahmathullah Syed (Norwich University), Jorge Talamantes (California State University – Bakersfield), Mark Timko (Elmhurst College), Paul Tjossem
provided a critical examination of the evolving manuscript and offered numerous suggestions for improvement—in addition to finding many previously undetected typographical glitches.

- especially Javier Hasbun (State University of West Georgia)—who had students working with almost the entire manuscript—but also William H. Ingham (James Madison University), Russel Kauffman (Muhlenberg College), Joelle L. Murray (Linfield College), Shafiq Rahman (Allegheny College), John Thompson (DePaul University), and Tim Young (University of North Dakota), all of whom used at least portions of the growing manuscript in their teaching during the academic year 2001–02 and who collectively sent me a wide spectrum of varied and valuable comments.

- several anonymous reviewers engaged by potential publishers as they evaluated my efforts, even though all potential publishers ultimately decided they could not provide the microscopic customization the book required.

Third, I wish to acknowledge considerable debt to many individuals whom I do not know but whose contributions behind the scenes have been invaluable. Chief among these individuals are:

- Donald Knuth, originator of \TeX;
- Leslie Lamport, originator of \LaTeX;
- Han The Than, originator of pdf\TeX, which provides the engine underlying \pdflatex for creating PDF files directly from \LaTeX source code;
- Pehong Chen and Nelson Beebe, originator and current maintainer of makeindex for preparing indices to \LaTeX documents;
- Enrico Gregorio, originator and current maintainer of \imakeidx that reduces the number of passes required to format and include an index;
- David Carlisle and Sebastian Rahtz, originators, and David Carlisle, current maintainer, of the \LaTeX package graphicx for including graphic images in \LaTeX documents;
- Leslie Lamport, David Carlisle and other members of the \LaTeX3 team, authors of the \LaTeX package ifthen for supporting conditional statements in \LaTeX source files;
- Sebastian Rahtz and Heiko Oberdiek, originator of the \LaTeX package hyperref for creating linked versions of documents as PDF files;
- Paul Vojta, author of \xdvi, a versatile on-screen previewer for the .dvi files produced by \TeX and \LaTeX;
- Till Tantau et. al., originators and current maintainers of the \LaTeX packages \tikz and \pgf for creating sophisticated diagrams using \LaTeX-like inclusions in the source file;
- William Chia-Wei Cheng, author of TGIF, a program used to create several of the figures;
- John Bradley, author of \xv, a program used to convert a few bit-mapped files into PostScript;
- Paul Bunson (Lawrence University), Albert Chen (Oklahoma Baptist University), K. Kelvin Cheng (Texas Tech University), Bob Davis (Taylor University), Mirela Fetea (University of Richmond), David Grumbine (St. Vincent College), Kristi Hendrickson LU ’91 (University of Puget Sound), Tom Herrmann (Eastern Oregon University), Seamus Lagan (Whittier College), Eric Lane (University of Tennessee at Chattanooga), Ernest Ma (Montclair State University), Norris Preyer (College of Charleston), Michael Schillaci (Francis Marion University), Mihir Sejpal (Earlham College), Laura Van Wormer (Hiram College), Mary Lou West (Montclair State University), and Douglas Young (Mercer University). In July, 2003: Stephen Addison (University of Central Arkansas), Mohan Aggarwal (Alabama A&M), Anne Caraley (SUNY at Oswego), William Dieterle (California University of Pennsylvania), Timothy Duman (University of Indianapolis), Brett Fadem (Colby College), Michael Gray (American University), James Meyer (St. Gregory’s University), Prabasaj Paul (Denison University), David Peterson (Francis Marion University), Edward Pogozelski (SUNY at Geneseo), Chandra Prayaga (University of West Florida), Bruce Richards (Oberlin College), David Seely (Albion College), Adam Smith (Hillsdale College), Walther Spjeldvik (Weber State University), Julie Talbot (State University of West Georgia), Paul Thomas (University of Wisconsin–Eau Claire), Srinivasa Venugopalan (SUNY at Binghamton), and Alma Zook (Pomona College).
• The many contributors to the MiKTeX project;
• The developers, authors, and maintainers of WinEdt, the text editor used during the latter years of writing CPSUP; and
• The developers and maintainers of ps2pdf for converting PostScript files to PDF and the developers and maintainers of pdfcrop for pruning excessive white borders from PDF files.
• Radical Eye Software, which holds the copyright on dvips, a program for converting .dvi files to PostScript.

Quite simply, this project would have been impossible without the availability of these several programs and utilities, each of which played a necessary role behind the scenes in preparing or processing the files from which, ultimately, a printable PostScript or PDF file for the finished book emerged.

Fourth, I point out that the names of several pieces of commercial software are, in fact, trademarks or registered trademarks belonging to the vendors of those software products. Each such trademark is identified at its first occurrence in the text proper, and detailed contact information for every vendor is compiled in Appendix Z.

Fifth, I acknowledge the following specific permissions, each of which is more fully explained at the point in the text where the permission is explicitly invoked. In particular, I thank

• The MathWorks, Inc., for permission to incorporate in this book and distribute IDL source code for the routines ludiffeq_23 and ludiffeq_45, which code uses algorithms patterned after those used in 1991 in the MATLAB routines ode23 and ode45.
• Wayne Landsman, author of the IDL routines qsimpson and trapzd in the IDL Astronomy User’s Library, for permission to use those routines as the basis for the routines luqsimp and lutrapzd and to distribute the source code for luqsimp and lutrapzd as supplements to this book.
• Research Systems (later Exelis Visual Information Solutions and now part of Harris Geospatial Solutions), Incorporated, for permission to use portions of any RSI-supplied and/or edited .pro code—most particularly evident in RSI contributions to ludiffeq_23.pro, ludiffeq_45.pro, and luqsimp.pro—and to use the IDL name and trademark.
• Numerical Recipes Software (a) for permission to use the names and calling sequences of several Numerical Recipes routines2 at various places in this book, (b) for permission to refer to the C header files nr.h and nrutil.h and the file nrutil.c containing assorted utilities used by various C recipes, and (3) for permission to use the names and calling sequences of several IDL routines that are derived from Numerical Recipes routines (and for the use of which Research Systems Incorporated has permission from Numerical Recipes Software).
• William Chia-Wei Cheng, author of TGIF, for permission to reproduce in the appendix on that program several of the icons used in its many screen displays.

Sixth, I wish to acknowledge and thank the W. M. Keck Foundation for three grants [#880969 (1988), #931248 (1993), #011795 (2001)], the National Science Foundation for three grants [USE-8851685 (ILI-1988), DUE-9350667 (ILI-1993), and DUE-9952285 (CCLI-EMD-2000)], and Lawrence University for matching and other support.3 A small conference held at Lawrence in 1990 and supported by the Alfred P. Sloan Foundation stimulated thinking among the participants about

---

2Specifically f1moon, xflmoon, caldat, julday, xjulday, avevar, xavevar, rk4, xrk4, rkqs, rkck, mmid, bsstep, rkdumb, odeint, trapzd, xtrapzd, qtrap, xqtrap, qsimp, qromb, point, rtbis, xrtbis, rtnext, xrtnewt, rtsafe, xrtsafe, zbrak, gaussj, ludcmp, lubksb, tridag, svdcmp, svbsv, mnewt, newt, and bryzdn (both in FORTRAN and in C).

3Any opinions, findings, and conclusions or recommendations expressed in this book are those of the author and do not necessarily reflect the views of any of these granting foundations or agencies.
uses of computers in upper-division undergraduate physics. All of these grants have contributed in many ways to the developments at Lawrence that have culminated in the writing of this book. In particular, the NSF CCLI-EMD grant made in February, 2000, supported my sabbatical while I finalized the text of (the first edition of) this book. That grant also supported four week-long summer faculty workshops that have, on the one hand, provided constructive feedback on a succession of drafts and, on the other hand, enhanced awareness nationally of this book and of the developments at Lawrence.

Finally (and most important of all), I wish to express deep gratitude for the decades during which my wife, Cynthia, has shown patience, understanding, and support beyond any possibility of my ever fully knowing. In truth, I do not know how to thank her.
Disclaimer

The statements described in the various chapters of this book have been tested extensively but have certainly not been tested with all versions of all software packages on all possible platforms with all possible versions of the underlying operating systems. Differences from version to version of the software packages, from operating system to operating system, and from platform to platform exist. This brief section identifies the versions of the various programs that have been tested and the operating systems and platforms on which those tests have been carried out. That the behavior of other combinations of version, operating system, and platform will conform in every detail to that herein described can, of course, not be guaranteed. One can, however, have some confidence that the behavior in combinations not explicitly tested will not differ enormously from that described herein—except that newer versions of a software package may well have features not implemented in earlier versions (and occasionally a feature or specific syntax available in an earlier version has been removed altogether from more recent versions). With reasonable confidence, one can presume that the commands and syntax and features described in this book will work on other platforms with the tested versions of the programs and with subsequent versions. Statements herein that exploit features implemented for the first time in the tested versions will, of course, not be accepted in earlier versions, but those “glitches” should not be numerous or extensive. Where, in the months and years since the original draft was created, I have become aware of such glitches, I have made appropriate updates and subsequent productions have incorporated those updates.\footnote{The date of production of each version of CPSUP is displayed at the top of the cover page on that version. I have maintained a dated list of edits made to the source files, so changes made after the date of production of a particular version of CPSUP and a subsequent production of that version can readily be identified for anyone who wishes to update an outdated production. Generally, updated productions fairly promptly replace the previous production at \url{parc.aapt.org/curricula/cpsup}. Versions dated between 10 and 31 January 2021 provide the base. Edits made after 31 January 2021 are recorded in the file of edits.} Nothing, however, assures that I have identified all such glitches.

That disclaimer having been stated, I now present for each program a brief tally of the version(s) tested and the platform(s) and operating system(s) on which those tests have been carried out:

- The MAXIMA codings herein have been fully verified with
  - MAXIMA Version 5.36.1 and wxMAXIMA Version 15.04.0 on a Hewlett-Packard platform running Windows 7,
  - MAXIMA Version 5.38.1 and wxMAXIMA Version 16.04.2 on a Hewlett-Packard platform running Windows 7, and
  - MAXIMA Version 5.36.1 and wxMAXIMA Version 12.01.0 on a Hewlett-Packard platform running the Fedora 17 implementation of LINUX.

In addition, these codings have been spot-checked with

- MAXIMA Version 5.38.0 and wxMAXIMA 16.04.1,
- MAXIMA 5.39.0 and wxMAXIMA 16.12.0, and
• MAXIMA 5.42.0 and wxMAXIMA 18.10.1

on a Hewlett-Packard platform running Windows 10.

• The MAPLE codings herein have been fully verified with MAPLE Version 16 on a Hewlett-Packard platform running Windows 7 and on a Hewlett-Packard platform running the Fedora 17 implementation of LINUX. These codings have also been spot-checked with MAPLE Version 17 on a Hewlett-Packard platform running Windows 10.

• The Mathematica codings have been fully verified with Mathematica Version 11.3 on a Hewlett-Packard platform running Windows 7 and on a Hewlett-Packard platform running the Fedora 17 implementation of LINUX. The Mathematica codings have also been spot-checked with Mathematica 12.0 on a Hewlett-Packard platform running Windows 10.

• The IDL details codings have been fully verified with IDL Versions 8.3 and 8.5 on a Hewlett-Packard platform running Windows 7 and a Hewlett-Packard platform running the Fedora 17 implementation of LINUX. These versions of IDL have also been spot-checked on a Hewlett-Packard platform running Windows 10.

• The MATLAB codings have been fully verified with MATLAB Version R2012a on a Hewlett-Packard Platform running Windows 7 and a Hewlett-Packard platform running the Fedora 17 implementation of LINUX. These codings have also been spot-checked on a Hewlett-Packard platform running Windows 10.

• The OCTAVE codings have been fully verified with

  o OCTAVE Version 4.0.0 on a Hewlett-Packard platform running Windows 7,
  o OCTAVE Version 3.6.3 on a Hewlett-Packard platform running the Fedora 17 implementation of LINUX, and
  o OCTAVE Version 4.0.3 on a Hewlett-Packard platform running the Fedora 25 implementation of LINUX.

OCTAVE Version 4.2.2 has been spot-checked on a Hewlett-Packard platform running Windows 7, and OCTAVE Versions 4.0.0 and 5.2.0 have been spot-checked on a Hewlett-Packard platform running Windows 10.

• Except where otherwise noted in the text, the PYTHON codings have been fully verified with

  o PYTHON 2.7.16 from the Anaconda2 distribution installed on a Hewlett-Packard platform running Windows 10, using the Anaconda2 prompt and also using the Anaconda2 Python Shell.  
  o PYTHON 3.7.3 from the Anaconda3 distribution installed on a Hewlett-Packard platform running Windows 10, using the Anaconda3 prompt and also using the Anaconda3 Python Shell.  

The codings in the PYTHON chapter and some of the PYTHON codings in other chapters have been verified with PYTHON 3.9.13 from the Anaconda3 distribution installed on a Hewlett-Packard platform running Windows 10, using the Anaconda3 Python shell.

• The Numerical Recipes codings have been fully verified with Numerical Recipes Version 2.10 only on a Hewlett-Packard platform running the Fedora 17 implementation of LINUX. Those codings have also been spot-checked on a Hewlett-Packard platform running Windows 10 using 64-bit GNU Fortran Version 7.1.0 and 64-bit GNU C Version 7.1.0.

---

\[2\] See the Local Guide for ways to bring up the prompt and the shell in your environment.

\[3\] See the previous footnote.
• The LSODE codings have been fully verified with LSODE whose README file bears the date 30 March 1987 only on a Hewlett-Packard platform running the Fedora-17 implementation of LINUX. The codings have also been spot-checked with LSODE whose README file (opkd-sum) bears the date 20 June 2001 on a Hewlett-Packard platform running Windows 10 with 64-bit GNU Fortran Version 7.1.0.

• The \LaTeX details apply specifically to \LaTeX 2\epsilon with the MiK\TeX implementation on a Hewlett-Packard platform running Windows 7 and on a Hewlett-Packard platform running Windows 10. \LaTeX normally responds to the same source code on all platforms.

• The TGIF details apply specifically to TGIF Version 4.2 (patchlevel 5) on a Hewlett-Packard platform running the Fedora 17 implementation of LINUX. (TGIF is exclusively a UNIX package.) Other versions and patchlevels and other platforms will surely have similar behavior, but may not conform exactly to the behavior here described.

• The MUDPACK codings have been fully verified with MUDPACK Version 5.0.1 whose README file bears the date 6 December 2011 only on a Hewlett-Packard platform running Windows 10 with 64-bit GNU Fortran Version 7.1.0.
# Table of Contents

Preface to the Second Edition ........................................ iii  
Acknowledgements ...................................................... vii  
Disclaimer ........................................................................ xi  
Table of Contents ......................................................... xv

## 1 Preliminaries
1.1 An Orientation to Computers ........................................ 2  
1.1.1 A Simple Responsive Machine ............................ 3  
1.1.2 Character Codes .............................................. 4  
1.1.3 The ASCII Character Set .................................. 5  
1.1.4 Representation of Data in a Computer ..................... 6  
1.2 Files and Directories ................................................. 8  
1.3 Operating Systems ....................................................... 9  
1.4 Glossary, Conventions, and Understandings .................. 9  
1.5 Assumed Background ................................................. 11  
1.5.1 The Gamma Function ..................................... 12  
1.5.2 The Laplace Transform .................................. 13  
1.6 Licensing Issues ..................................................... 14

## 4 Introduction to OCTAVE  
4.1 Beginning an OCTAVE Session .................................. 16  
4.2 Basic Entities in OCTAVE ......................................... 17  
4.2.1 Data Types .................................................. 17  
4.2.2 Variable Names .............................................. 17  
4.2.3 Assignment of Values to Variables ....................... 18  
4.2.4 Commands .................................................... 18  
4.3 A Sampling of OCTAVE Capabilities ......................... 19  
4.3.1 Creating and Examining Arrays ......................... 19  
4.3.2 Simple Mathematical Manipulations ..................... 22  
4.3.3 Matrix Manipulations .................................... 23  
4.3.4 Solving Linear Equations ................................ 24  
4.3.5 A First Graph .............................................. 25  
4.4 Properties, Objects, and Handles ......................... 28  
4.5 Saving and Retrieving the OCTAVE Session ............... 30  
4.6 Loops, Logical Expressions, and Conditionals ............. 32  
4.7 Reading Data from a File .................................... 35  
4.8 On-line Help ....................................................... 37  
4.9 m-Files .......................................................... 37  
4.10 Eigenvalues and Eigenvectors ................................. 41  
4.11 Graphing Scalar Functions of One Variable ............... 43  
4.11.1 The Basic Strategy ..................................... 43  
4.11.2 Marking Points: .......................................... 43  
4.11.3 Plotting Several Graphs on One Set of Axes ......... 44  
4.11.4 Polar Plots and Custom Axes ............................ 45
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.11.5</td>
<td>Multiple Plots On a Page</td>
</tr>
<tr>
<td>4.11.6</td>
<td>Plotting Experimental Data</td>
</tr>
<tr>
<td>4.12</td>
<td>Making Hard Copy</td>
</tr>
<tr>
<td>4.12.1</td>
<td>... of Text</td>
</tr>
<tr>
<td>4.12.2</td>
<td>... of Graphs</td>
</tr>
<tr>
<td>4.13</td>
<td>Graphing Scalar Functions of Two Variables</td>
</tr>
<tr>
<td>4.13.1</td>
<td>A Preliminary: The Function <code>meshgrid</code></td>
</tr>
<tr>
<td>4.13.2</td>
<td>Surface Plots: The Function <code>mesh</code></td>
</tr>
<tr>
<td>4.13.3</td>
<td>Shaded Surfaces: The Function <code>surf</code></td>
</tr>
<tr>
<td>4.13.4</td>
<td>Contour Plots: The Function <code>contour</code></td>
</tr>
<tr>
<td>4.13.5</td>
<td>The Functions <code>meshc</code>, <code>surf</code>, and <code>surfl</code></td>
</tr>
<tr>
<td>4.13.6</td>
<td>Functions of Two Variables in Polar Coordinates</td>
</tr>
<tr>
<td>4.14</td>
<td>Graphing Scalar Functions of Three Variables</td>
</tr>
<tr>
<td>4.14.1</td>
<td>Reduction to Two-Dimensional Displays</td>
</tr>
<tr>
<td>4.14.2</td>
<td>Isosurfaces</td>
</tr>
<tr>
<td>4.15</td>
<td>Graphing Vector Fields</td>
</tr>
<tr>
<td>4.15.1</td>
<td>The Function <code>quiver</code></td>
</tr>
<tr>
<td>4.15.2</td>
<td>More Elaborate Displays</td>
</tr>
<tr>
<td>4.15.3</td>
<td>Three-Dimensional Vector Fields</td>
</tr>
<tr>
<td>4.16</td>
<td>Animation</td>
</tr>
<tr>
<td>4.17</td>
<td>Advanced Graphing Features</td>
</tr>
<tr>
<td>4.17.1</td>
<td>Fonts</td>
</tr>
<tr>
<td>4.17.2</td>
<td>Space Curves</td>
</tr>
<tr>
<td>4.17.3</td>
<td>Using Multiple Windows</td>
</tr>
<tr>
<td>4.17.4</td>
<td>Customizing Axes</td>
</tr>
<tr>
<td>4.18</td>
<td>Miscellaneous Occasionally Useful Tidbits</td>
</tr>
<tr>
<td>4.18.1</td>
<td>Specifying Directories</td>
</tr>
<tr>
<td>4.18.2</td>
<td>Deleting Variables No Longer Needed</td>
</tr>
<tr>
<td>4.18.3</td>
<td>OCTAVE's Current Directory</td>
</tr>
<tr>
<td>4.18.4</td>
<td>OCTAVE's Search Path</td>
</tr>
<tr>
<td>4.18.5</td>
<td>Customizing OCTAVE</td>
</tr>
<tr>
<td>4.18.6</td>
<td>Restoring OCTAVE's Initial State</td>
</tr>
<tr>
<td>4.18.7</td>
<td>The Command <code>format</code></td>
</tr>
<tr>
<td>4.18.8</td>
<td>Reading Data from the Keyboard</td>
</tr>
<tr>
<td>4.18.9</td>
<td>Writing Data to the Screen</td>
</tr>
<tr>
<td>4.18.10</td>
<td>Exiting from Procedures</td>
</tr>
<tr>
<td>4.19</td>
<td>References</td>
</tr>
<tr>
<td>4.20</td>
<td>Exercises</td>
</tr>
<tr>
<td>4.20.1</td>
<td>Writing OCTAVE Statements</td>
</tr>
<tr>
<td>4.20.2</td>
<td>Finding Eigenvalues and Eigenvectors</td>
</tr>
<tr>
<td>4.20.3</td>
<td>Graphing Scalar Functions of a Single Variable</td>
</tr>
<tr>
<td>4.20.4</td>
<td>Graphing Scalar Functions of Two Variables</td>
</tr>
<tr>
<td>4.20.5</td>
<td>Graphing Scalar Functions of Three Variables</td>
</tr>
<tr>
<td>4.20.6</td>
<td>Graphing Two-Dimensional Vector Fields</td>
</tr>
<tr>
<td>4.20.7</td>
<td>Animation</td>
</tr>
<tr>
<td>4.20.8</td>
<td>Supplementary Exercises</td>
</tr>
</tbody>
</table>

8 Introduction to **MATHEMATICA**

8.1 Beginning a *Mathematica* Session | 92 |
8.2 On-Line Help | 93 |
8.3 Basic Entities in *Mathematica* | 94 |
8.4 Variable Names in *Mathematica* | 94 |
8.5 Expressions in *Mathematica* | 95 |
8.6 Assigning Values to Variables; Defining Functions | 96 |
8.7 Creating and Examining Lists | 99 |
8.8 Using *Mathematica* | 100 |
8.8.1 ... for Arithmetic | 101 |
8.8.2 ... for Algebra | 102 |
8.8.3 ... for Complex Variables | 105 |
8.8.4 ... for Trigonometry | 107 |
TABLE OF CONTENTS

8.8.5 ... for Algebraic Equations .......................................................... 108
8.8.6 ... for Manipulating Vectors and Matrices .................................... 110
8.8.7 ... for Basic Calculus .................................................................... 111
8.8.8 ... for Laplace Transforms .............................................................. 114
8.8.9 ... for Ordinary Differential Equations ........................................... 115
8.8.10 ... for Vector Calculus ................................................................. 119
8.9 Packages ......................................................................................... 121
8.10 Loops, Logical Expressions, and Conditionals .................................... 122
8.11 Command Files ................................................................................ 124
8.12 High-Resolution Graphing ................................................................. 126
8.13 Making Hard Copy ........................................................................... 129
  8.13.1 ... from the NBI ........................................................................ 130
  8.13.2 ... from the TBI ........................................................................ 132
8.14 Output in \LaTeX\ Format .................................................................. 132
8.15 Animation ....................................................................................... 133
8.16 Using the Notebook .......................................................................... 134
8.17 Customizations ............................................................................... 135
8.18 Miscellaneous Occasionally Useful Tidbits ...................................... 136
  8.18.1 Turning off Messages .................................................................. 136
  8.18.2 Specifying Directories .................................................................. 136
  8.18.3 Default Directory ...................................................................... 136
  8.18.4 Search Path ................................................................................ 137
  8.18.5 Displaying and Changing Default Options .................................... 137
  8.18.6 Placing Text in Graphs ................................................................. 138
  8.18.7 Space Curves ............................................................................. 138
  8.18.8 Directing Plot to Separate Window ............................................... 139
8.19 References ..................................................................................... 140
8.20 Exercises ....................................................................................... 140

9 Introduction to Programming ............................................................. 151
  9.1 Components of a Programming Language ....................................... 152
    9.1.1 Variables, Variable Names, and Use of Memory ......................... 152
    9.1.2 Essential Action Statements ....................................................... 154
    9.1.3 Logical Expressions and Conditions ............................................ 155
    9.1.4 Syntactic Wrinkles ................................................................... 156
    9.1.5 Essential Control Structures ....................................................... 156
  9.2 Sample Short Algorithms .................................................................. 160
  9.3 Two Larger Algorithms .................................................................... 162
    9.3.1 Solving Laplace’s Equation ......................................................... 162
    9.3.2 File Output/Input ..................................................................... 164
  9.10 Solving Laplace’s Equation with OCTAVE ..................................... 167
  9.11 Making Hard Copy .......................................................................... 169
  9.12 Creating and Storing Two-Dimensional Scalar Arrays .................... 169
    9.12.5 ... with OCTAVE ................................................................. 169
  9.13 Creating and Storing Three-Dimensional Scalar Arrays .................. 170
    9.13.5 ... with OCTAVE ................................................................. 170
  9.14 Creating and Storing Two-Dimensional Vector Arrays .................... 170
    9.14.5 ... with OCTAVE ................................................................. 170
  9.15 Creating and Storing Three-Dimensional Vector Arrays .................. 172
    9.15.5 ... with OCTAVE ................................................................. 172
  9.16 Reading Files .................................................................................. 174
    9.16.5 ... with OCTAVE ................................................................. 174
  9.17 References ..................................................................................... 175
  9.18 Exercises ....................................................................................... 175
  9.G Listing of \texttt{lugen\_list\_m (OCTAVE)} ........................................ 181
  9.S Listing of \texttt{ludead\_list\_m (OCTAVE)} ........................................ 183

11 Solving Ordinary Differential Equations ............................................ 185
  11.1 Sample Problems ........................................................................... 186
    11.1.1 Projectile in a Viscous Medium ................................................. 186
    11.1.2 Chain Radioactive Decay ......................................................... 189
# TABLE OF CONTENTS

11.1.3 Exponential and Logistic Growth ........................................... 190
11.1.4 Forced, Driven, Damped Harmonic Oscillation ............................... 190
11.1.5 An LRC Resonant Circuit .................................................. 192
11.1.6 Coupled Oscillators ...................................................... 192
11.1.7 Motion under Central Forces ............................................. 193
11.1.8 Standing Waves in a String .............................................. 195
11.1.9 The Schrödinger Equation in One Dimension ............................... 196
11.2 Laplace Transforms .......................................................... 197
11.5 Solving ODEs Symbolically with Mathematica .................................. 198
  11.5.1 Projectile in a Viscous Medium ......................................... 199
  11.5.2 Logistic Growth .......................................................... 201
  11.5.3 Damped Harmonic Oscillator ........................................... 202
  11.5.4 Chain Radioactive Decay ............................................... 202
  11.5.5 Coupled Oscillators .................................................... 205
  11.5.6 Standing Waves in a String ............................................ 206
  11.5.7 Infinite Depth Quantum Well .......................................... 209
  11.5.8 Finding Solutions by Series Methods ................................ 209
11.6 Algorithms for Solving ODEs Numerically .................................... 212
  11.6.1 Euler’s Method ................................................................ 212
  11.6.2 Improved Euler Method .................................................. 214
  11.6.3 Runge-Kutta Methods .................................................... 216
  11.6.4 Assessing Accuracy ....................................................... 218
  11.6.5 Adaptive Methods ....................................................... 219
  11.6.6 Multistep Methods ....................................................... 220
11.9 Solving ODEs Numerically with OCTAVE ...................................... 221
  11.9.1 Using Elementary Commands ........................................... 221
  11.9.2 Defining ODEs for OCTAVE ............................................ 224
  11.9.3 The Procedure lsode ..................................................... 225
  11.9.4 Chain Radioactive Decay ............................................... 226
  11.9.5 Damped Harmonic Oscillator ........................................... 229
  11.9.6 Planetary Orbits .......................................................... 230
  11.9.7 Standing Waves in a String ............................................ 232
  11.9.8 The Quantum Harmonic Oscillator ................................... 237
  11.9.9 Changing Lsode Options ............................................... 242
  11.9.10 M-files in the Public Library ........................................ 244
11.13 Solving ODEs Numerically with Mathematica ................................ 245
11.16 Exercises ............................................................................... 247
  11.16.1 Using Symbolic Methods ............................................... 247
  11.16.2 Using Numerical Methods ............................................... 250
13 Evaluating Integrals ...................................................................... 255
  13.1 Sample Problems ................................................................... 255
    13.1.1 One-Dimensional Trajectories .......................................... 255
    13.1.2 Center of Mass ............................................................ 256
    13.1.3 Moment of Inertia; Radius of Gyration ................................. 257
    13.1.4 The Large Amplitude Simple Pendulum ............................... 258
    13.1.5 Statistical Data Analysis ............................................... 259
    13.1.6 The Cornu Spiral ....................................................... 260
    13.1.7 Electric and Magnetic Fields and Potentials .......................... 260
    13.1.8 Quantum Probabilities ............................................... 262
    13.1.9 Expansion in Orthogonal Functions .................................. 262
  13.4 Evaluating Integrals Symbolically with Mathematica ....................... 263
    13.4.1 Relativistic Motion Under a Constant Force .......................... 264
    13.4.2 Center of Mass ............................................................ 265
    13.4.3 Moment of Inertia; Radius of Gyration ................................ 266
    13.4.4 Electrostatic Potential of a Finite Line Charge ........................ 266
    13.4.5 The Helmholtz Coil ..................................................... 268
    13.4.6 Period of a Pendulum: Correction as Amplitude Grows ............ 272
    13.4.7 Fourier Coefficients for Half-Rectified Signal ....................... 273
  13.5 Algorithms for Numerical Integration ......................................... 276
TABLE OF CONTENTS  

13.5.1 Newton-Cotes Quadrature ........................................ 276
13.5.2 Rearrangements for Computational Efficiency .................. 277
13.5.3 Assessing Error .................................................... 279
13.5.4 Iterative and Adaptive Algorithms .............................. 281
13.5.5 Gaussian Quadrature .............................................. 281
13.8 Evaluating Integrals Numerically with OCTAVE .................... 284
  13.8.1 Using Elementary Commands .................................... 284
  13.8.2 Built-In Integration Routines .................................. 287
  13.8.3 Moment of Inertia ............................................... 290
  13.8.4 Quantum Probabilities ......................................... 291
  13.8.5 Integrals as Functions of the Upper Limit .................. 292
  13.8.6 The Error Function ............................................. 292
  13.8.7 The Cornu Spiral ............................................... 292
  13.8.8 Integrals as Functions of an Internal Parameter .......... 294
  13.8.9 The Off-Axis Electrostatic Potential of Two Rings ....... 295
13.12 Evaluating Integrals Numerically with Mathematica .............. 296
  13.12.1 Quantum Probability ........................................... 298
  13.12.2 The Error Function ............................................ 298
  13.12.3 The Off-Axis Electrostatic Potential of Two Rings ....... 299
13.15 Exercises ............................................................ 300
  13.15.1.1 using Symbolic Methods ................................. 300
  13.15.2.1 using Numerical Methods ............................... 303
14 Finding Roots ................... 307
  14.1 Sample Problems .................................................... 307
    14.1.1 Classical Turning Points .................................... 308
    14.1.2 A Max-Min Problem: Equilibrium Points .................. 308
    14.1.3 Natural Frequencies of Oscillating Systems ............ 309
    14.1.4 Range of Projectile in Viscous Medium ................ 310
    14.1.5 Energy Levels in a Quantum Well ....................... 313
  14.2 Symbolic Approaches ............................................... 314
  14.5 Finding Roots Symbolically with Mathematica .................... 315
    14.5.1 Classical Turning Points .................................... 316
    14.5.2 Equilibrium Points ......................................... 316
    14.5.3 Natural Frequencies ........................................ 317
    14.5.4 Range of Projectile in Viscous Medium ............... 318
  14.6 Algorithms for Finding Roots Numerically ...................... 320
    14.6.1 The Method of Bisection .................................... 320
    14.6.2 Newton’s Method ............................................. 321
    14.6.3 Other Methods ................................................ 321
    14.6.4 Assessing Accuracy .......................................... 322
  14.9 Finding Roots Numerically with OCTAVE ........................... 323
    14.9.1 Classical Turning Points .................................... 325
    14.9.2 Natural Frequencies ........................................ 326
    14.9.3 Range of Projectile ......................................... 327
    14.9.4 Energy Levels in Quantum Well ............................ 330
  14.13 Finding Roots Numerically with Mathematica .................... 332
    14.13.1 Classical Turning Points .................................... 335
    14.13.2 Range of Projectile ......................................... 335
    14.13.3 Energy Levels in Quantum Well ............................ 336
  14.16 Solving Simultaneous Equations ................................ 339
    14.16.1 Systems of Linear Equations ............................... 340
    14.16.2 Systems of Nonlinear Equations ........................... 343
  14.17 Exercises ............................................................ 345
    14.17.1.1 using Symbolic Methods ................................ 345
    14.17.2.1 using Numerical Methods ............................. 347
    14.17.4 Finding More than One Unknown ............................ 349
  14.C Listing of OCTAVE Function lubisect.m .......................... 354

15 Solving Partial Differential Equations ............................ 355
# TABLE OF CONTENTS

15.1 Sample Problems .................................................. 357
  15.1.1 Motion of a String: The Wave Equation .................... 357
  15.1.2 Heat Flow: The Diffusion Equation ....................... 360
  15.1.3 Steady State Heat Flow: The Laplace Equation ............. 361
  15.1.4 Other Situations ........................................... 361
  15.1.5 Classification of Second-Order PDEs ...................... 362
  15.1.6 Equations of Fluid Dynamics: Navier-Stokes' Equations .... 363
    15.1.6.1 Conservation of Mass: The Equation of Continuity .. 364
    15.1.6.2 Newton's Second Law ................................ 365
    15.1.6.3 A Special Case: Non-viscous Flow ................... 367
    15.1.6.4 A Second Special Case: Sound Waves ................. 368
  15.1.7 A General 1D Equation .................................... 369
  15.1.8 A General 2D Equation .................................... 370
  15.2 Finite Difference Methods (FDMs) in One Dimension .......... 370
  15.5 Using OCTAVE to Solve 1D PDEs via an FDM ................. 373
    15.5.1 A General Coding ...................................... 373
    15.5.2 An Example: Simple Harmonic Motion .................... 374
  15.9 Finite Element Methods (FEMs) in One Dimension .......... 377
    15.9.1 Discretizing the Domain: Pre-processing ............... 378
    15.9.2 Selecting Interpolation or Shape Functions ............. 378
    15.9.3 Formulating the Equations for a Single Element ......... 379
    15.9.4 Assembling the System of Equations .................... 382
    15.9.5 Incorporating the Boundary Conditions ................. 384
  15.12 Using OCTAVE to Solve 1D PDEs via an FEM ............... 386
    15.12.1 A General Coding ...................................... 386
    15.12.2 An Example: Simple Harmonic Motion .................... 388
  15.16 Finite Difference Methods (FDMs) in Two Dimensions .... 391
    15.16.1 The Wave Equation ..................................... 391
    15.16.2 The Diffusion Equation ................................ 393
    15.16.3 The Laplace Equation ................................ 396
  15.19 Using OCTAVE to Solve 2D PDEs via an FDM ............... 399
    15.19.1 The Wave Equation ..................................... 399
    15.19.2 The Diffusion Equation ................................ 404
    15.19.3 The Laplace Equation ................................ 410
  15.23 Finite Element Methods (FEMs) in Two Dimensions ........ 416
    15.23.1 Discretizing the Domain: Pre-processing ............... 417
    15.23.2 Selecting the Interpolation or Shape Functions ......... 419
    15.23.3 Formulating the Equations for a Single Element ......... 420
    15.23.4 Assembling the System of Equations .................... 422
    15.23.5 Incorporating the Boundary Conditions ................. 424
  15.26 Using OCTAVE to Solve 2D PDEs via an FEM ............... 427
    15.26.1 A General Coding ...................................... 427
    15.26.2 An Example: Isotropic Heat Flow ....................... 434
  15.30 Exercises .................................................... 437
  15.A Program(s) for FDM Approach to 1D Problem ................ 445
    15.A.3 Listing of fdmd.m (OCTAVE) ............................ 445
  15.B Evaluating Integrals for 1D FEM Problem .................... 446
    15.B.3 ... using Mathematica .................................. 446
  15.C Program(s) for FEM Approach to 1D Problem ................ 446
    15.C.3 Listing of femd.m (OCTAVE) ............................ 446
  15.D OCTAVE Programs for FDM Approach to 2D Problems .......... 448
    15.D.1 Listing of fmdwave1d.m ................................ 448
    15.D.2 Listing of fmdiffus1d.m ................................ 449
    15.D.3 Listing of fdm1d.m ....................................... 450
    15.D.4 Listing of fdm2d.pplot.m ............................... 451
  15.E A Useful Integral ............................................ 452
  15.F Evaluating Integrals for 2D FEM Problem .................... 453
    15.F.3 ... using Mathematica .................................. 453
  15.G Program(s) for FEM Approach to 2D Problems ................ 456
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.1</td>
<td>Creating a Simple Document</td>
<td>460</td>
</tr>
<tr>
<td>A.1.1</td>
<td>Structuring a \LaTeX{} Source File</td>
<td>460</td>
</tr>
<tr>
<td>A.1.2</td>
<td>Creating a PostScript File</td>
<td>462</td>
</tr>
<tr>
<td>A.1.3</td>
<td>Creating a PDF File</td>
<td>463</td>
</tr>
<tr>
<td>A.1.4</td>
<td>Displaying the Document on the Screen</td>
<td>464</td>
</tr>
<tr>
<td>A.1.5</td>
<td>Printing the Document</td>
<td>464</td>
</tr>
<tr>
<td>A.2</td>
<td>Specification of Global Style: The Preamble</td>
<td>465</td>
</tr>
<tr>
<td>A.3</td>
<td>In-Text Specification of Local Style</td>
<td>467</td>
</tr>
<tr>
<td>A.3.1</td>
<td>Type Style</td>
<td>467</td>
</tr>
<tr>
<td>A.3.2</td>
<td>Type Size</td>
<td>468</td>
</tr>
<tr>
<td>A.3.3</td>
<td>White Space</td>
<td>469</td>
</tr>
<tr>
<td>A.3.4</td>
<td>Environments</td>
<td>470</td>
</tr>
<tr>
<td>A.3.5</td>
<td>\LaTeX{} Packages</td>
<td>471</td>
</tr>
<tr>
<td>A.3.6</td>
<td>Miscellaneous Other Capabilities</td>
<td>471</td>
</tr>
<tr>
<td>A.4</td>
<td>Including Equations</td>
<td>472</td>
</tr>
<tr>
<td>A.5</td>
<td>Including Lists</td>
<td>473</td>
</tr>
<tr>
<td>A.6</td>
<td>Including Tables</td>
<td>474</td>
</tr>
<tr>
<td>A.7</td>
<td>Including Illustrations</td>
<td>476</td>
</tr>
<tr>
<td>A.7.1</td>
<td>Using Cut and Paste</td>
<td>476</td>
</tr>
<tr>
<td>A.7.2</td>
<td>Using the \texttt{graphicx} Package</td>
<td>477</td>
</tr>
<tr>
<td>A.7.3</td>
<td>Using the \texttt{tikz} Package</td>
<td>479</td>
</tr>
<tr>
<td>A.7.4</td>
<td>Using the \texttt{picture} Environment and \texttt{pict2e} Package</td>
<td>482</td>
</tr>
<tr>
<td>A.8</td>
<td>Including TOC, LOF, and LOT</td>
<td>482</td>
</tr>
<tr>
<td>A.9</td>
<td>Including an Index</td>
<td>483</td>
</tr>
<tr>
<td>A.10</td>
<td>Including Hyperlinks</td>
<td>485</td>
</tr>
<tr>
<td>A.11</td>
<td>Converting .eps and .ps Files to .pdf</td>
<td>486</td>
</tr>
<tr>
<td>A.12</td>
<td>Using Conditional Expressions in \LaTeX{}</td>
<td>488</td>
</tr>
<tr>
<td>A.13</td>
<td>Error Messages Generated by \LaTeX{}</td>
<td>489</td>
</tr>
<tr>
<td>A.14</td>
<td>The Page Previewer for .dvi Files</td>
<td>490</td>
</tr>
<tr>
<td>A.15</td>
<td>The Spell Checker in UNIX</td>
<td>491</td>
</tr>
<tr>
<td>A.16</td>
<td>A Sample Document</td>
<td>492</td>
</tr>
<tr>
<td>A.17</td>
<td>Miscellaneous Other Features</td>
<td>493</td>
</tr>
<tr>
<td>A.18</td>
<td>References</td>
<td>500</td>
</tr>
<tr>
<td>A.19</td>
<td>Exercises</td>
<td>501</td>
</tr>
<tr>
<td>A.A</td>
<td>Listings</td>
<td>504</td>
</tr>
<tr>
<td>A.A.1</td>
<td>for Windows</td>
<td>504</td>
</tr>
<tr>
<td>A.A.1.1</td>
<td>Batch File \texttt{ceps2pdf.bat}</td>
<td>504</td>
</tr>
<tr>
<td>A.A.1.2</td>
<td>Batch File \texttt{cps2pdf.bat}</td>
<td>504</td>
</tr>
<tr>
<td>A.A.1.3</td>
<td>Batch File \texttt{rdfileeps.bat}</td>
<td>504</td>
</tr>
<tr>
<td>A.A.1.4</td>
<td>Batch File \texttt{rdfileps.bat}</td>
<td>504</td>
</tr>
<tr>
<td>A.A.2</td>
<td>for UNIX</td>
<td>505</td>
</tr>
<tr>
<td>A.A.2.1</td>
<td>Shell Script \texttt{ceps2pdf}</td>
<td>505</td>
</tr>
<tr>
<td>A.A.2.2</td>
<td>Shell Script \texttt{cps2pdf}</td>
<td>505</td>
</tr>
<tr>
<td>A.A.2.3</td>
<td>Shell Script \texttt{rdfileps}</td>
<td>505</td>
</tr>
<tr>
<td>A.A.2.4</td>
<td>Shell Script \texttt{rdfileps}</td>
<td>505</td>
</tr>
<tr>
<td>A.A.3</td>
<td>for Windows and UNIX</td>
<td>506</td>
</tr>
<tr>
<td>A.A.3.1</td>
<td>Python Script \texttt{ExtractFileName.py}</td>
<td>506</td>
</tr>
</tbody>
</table>

Z Contacting Software Vendors 507

Index 511
Chapter 1

Preliminaries

Over the past two decades, acquaintance with computational approaches to problems—and with the computational resources that facilitate those approaches—has come to be critically important for success in the sciences. This book aims to develop familiarity with a variety of computational tools and techniques in application particularly to problems in physics. Rather than selecting a single application program, we presume that productive use of contemporary computational resources requires acquaintance with several different sorts of tools, including:

- an array processing program (e.g., IDL®, MATLAB®, OCTAVE, PYTHON, ...);
- a computer algebra system (e.g., MAPLE®, Mathematica®, MAXIMA, ...);
- a standard scientific programming language (e.g., FORTRAN, C, PYTHON, ...), both for programming ab initio and, more particularly, for creating driving programs to invoke commercially available subroutine packages like NUMERICAL RECIPES, and freely downloadable subroutine packages like LSODE, MUDPACK, and LAPACK; and
- a tool for graphical visualization of scalar and vector functions of one, two, and three independent variables (e.g., IDL, MATLAB, OCTAVE, PYTHON, MAXIMA, MAPLE, Mathematica, ...).

Further, to make effective use of these tools, the user must

- be acquainted with the main capabilities of at least one operating system (e.g., UNIX, Windows, Macintosh OS, ...),
- be fluent in the use of a text editor (e.g., gedit, xemacs, vim, winedt, ...) and of a program for creating drawings (e.g., tgif, ...), and
- of a publishing package (e.g., TeXlive, \LaTeX, MiKTeX, Oz\TeX, ...) capable of formatting elaborate equations, incorporating symbolic references within documents, generating tables of contents and indices, and so on.

This book introduces intermediate-level physics students to a selected spectrum of these tools, helps them learn enough of the tools' capabilities to know what the tools can do, and builds their confidence both in using the tools and in reading vendor-supplied documentation. Ultimately, we expect that

---

Many of the specific examples in this list are identified by names that are trademarks belonging to the vendor of the identified software and registered in the United States Patent and Trademark Office. Those that the author knows to have that status are identified with the symbol ® at the first occurrence of the name. Full contact information for the vendors of the software (and the owners of the trademarks) is compiled in Appendix Z.
students launched into the computational world as sophomores, say, will, as juniors and seniors, be motivated to use computational resources intelligently and successfully on their own initiative, whenever it seems to them appropriate to exploit those tools. As a resource, the computer should parallel the library; this book aims to help students develop the skills to support that view.\footnote{Uses of internet resources are conspicuously absent from the list of skills in this opening paragraph. While such uses are playing an increasingly important role both in education and in professional life, they are explicitly excluded from the purview of this book.}

In this book, the ultimate objective described in the previous paragraph is pursued in several steps:

1. You learn to manipulate the system you have and to work efficiently with whatever text editor is available. For the most part, this step is the task of the \textit{Local Guide}.

2. You learn the basic commands for one or more tools (how to start the tool, how to stop the tool, how to construct the primary entities—mathematical expressions, numerical arrays, ...—on which the tool works, how to manipulate those entities, how to generate output—both textual and graphical—from the tool, etc.). This step is the business of the first portion of this book and of the appendices.

3. You learn ways in which these tools can be used to advantage to address prototype problems in a variety of areas of physics. This step is the business of the second portion of this book, each chapter of which begins by describing several representative problems that involve a particular type of computation (solving ODEs, integrating, finding roots, ...). Then, each chapter addresses those problems with a succession of computational tools, some symbolic, some numeric—exploiting graphical displays whenever appropriate to the exercise at hand. Each chapter concludes with numerous exercises to direct your own further study of the tools and techniques addressed in the chapter.

You need not, of course, complete all of one step before proceeding to some of the next step. Once you have learned to manipulate your computer system and use an available text editor, you can pick and choose the tools and examples of greatest—or most immediate—interest to you. To be sure, some portions of earlier chapters are prerequisite to some portions of later chapters, but the linkages are neither deep nor extensive. Thus, you can hop around in this book as your needs and interests dictate.

In the remainder of this chapter, we address several general items relating to the design and use of computers and to the structure of this book. Here and there, specific items may well be site dependent. Thus, as a companion to this book, you must obtain from your local site administrator a copy of the \textit{Local Guide}, which supplements this book with detailed information that relates specifically to your site.

Be aware, in particular, that many of the chapters in this book are at least in part tutorial in nature. Full study of the material here presented requires you to replicate the illustrated “conversations” with the computer. To do so, you must—of course—be logged into an appropriate computer system, as described in the \textit{Local Guide}. This paragraph, however, is the only point in the book at which the wisdom of being logged in is explicitly mentioned.

\section{1.1 An Orientation to Computers}

We begin by inventing (at least some aspects of) a computer, in the process motivating some of its main features and discussing briefly a few important underlying concepts and structures.
1.1. A Simple Responsive Machine

Consider first a typewriter. In broad outline, its user commands the printing mechanism (hereafter printer) to perform a desired sequence of actions by pressing the corresponding sequence of keys on the keyboard. Most keys cause the printer to print a particular character on the paper and advance the printhead to its next position. When the key labeled ‘a’ is pressed, for example, the character ‘a’ is printed on the paper and the printhead is advanced; when the shift key is held down while the key labeled ‘5’ is pressed (sometimes denoted (SHIFT/5)), the character ‘%’ is printed and the printhead is advanced; etc. A few keys command the printer to perform other actions. Pressing the space bar, for example, advances the printhead without printing a visible character. (Actually, it is useful to think that the space character, denoted (SP), has been “printed”.) Pressing the key labeled RETURN “prints” the carriage return character (denoted ⟨CR⟩), which moves the printhead to the beginning of the line and advances or feeds the paper one line further along.

We can, however, imagine a more general “typewriter”—i.e., a computer—in which an obedient and instructable “agent”—hereafter the central processing unit (CPU)—has been interposed between the keyboard and the printer. Further, let us build this expanded machine so that (a) pressing a key at the keyboard sends a (probably electrical) code identifying that key to the CPU and (b) the printer interprets and responds to each code received from the CPU. This machine reverts to our original typewriter if we tell the CPU to carry out or execute the statements or commands.

```
LOOP
  Read code from keyboard
  Send code to printer
END_LOOP
```

The action of the machine in response to representative key strokes would then be described as follows:

- When the key labeled ‘a’ is pressed, the keyboard sends the code for the character ‘a’ to the CPU, which then transmits that code to the printer.
- When the shift key is held down while the key labeled ‘5’ is pressed, the keyboard sends the code for the character ‘%’ to the CPU, which then transmits that code to the printer.
- When the space bar is pressed, the keyboard sends the code for the character ⟨SP⟩ to the CPU, which then transmits that code to the printer.
- When the key labeled RETURN is pressed, the keyboard sends the code for the character ⟨CR⟩ to the CPU, which then transmits that code to the printer.

In the first three cases, the printer displays the character identified by the received code and also advances the printhead. In the fourth case, the printer should both return the printhead and feed the paper. In fact, most printers treat returning the printhead and feeding the paper as two distinct operations. Receipt of the code for the character (CR) will effect the former operation; receipt of a different code, that for the line-feed character (LF), will effect the latter. While it is convenient to have a single keystroke at the keyboard accomplish both operations, most printers must receive two separate codes to accomplish the desired action. Thus, we must tell the CPU that receipt of the code for the character (CR) from the keyboard must trigger the sending of the codes for the pair of characters (CR)/LF to the printer. To simulate a typewriter, we must embellish the above statements to

---

3The special words LOOP and END_LOOP bracket a group or block of instructions that are as a block to be executed repeatedly. We shall here ignore concerns about stopping the loop.

4The special words IF, THEN, and END_IF convey a conditional execution of one or more statements. The statement(s) between the THEN and the END_IF will be executed only if the condition following the IF is true when the entire construction is encountered.
PROGRAM TYPEWRITER

LOOP
  Read code from keyboard
  Send code to printer
  IF code is that for <CR>
    THEN Send code for <LF> to printer
  END_IF
END_LOOP

END_PROGRAM

In this listing, we have introduced the word program to identify a complete set of instructions for the performance of some task, and we have introduced the special words PROGRAM and END_PROGRAM to bracket a program. We have also provided a way to designate an appropriate name for the program.

Note that, while a particular code is always associated with a character, not all codes are associated with printing characters. Non-printing characters are called control characters. When received by a printer (or other peripheral device), they result not in the display of a particular symbol but in the performance of some other function. We have already met ⟨CR⟩ and ⟨LF⟩. Other control characters familiar to the user of an ordinary typewriter are the backspace ⟨BS⟩, which causes the printhead to back up one space; the horizontal tab ⟨HT⟩, which causes the printhead to advance to the next pre-set (horizontal) tab position; and the vertical tab ⟨VT⟩, which advances the paper to the next preset (vertical) tab position.

1.1.2 Character Codes

To facilitate visualizing the codes seen by the CPU, imagine that the CPU receives its signals by “looking at” a row of eight light bulbs. Further, declare that pressing a particular key on the keyboard turns some of the bulbs on and leaves the rest off, and endow the CPU with a capacity to sense which bulbs are on and which off. If we represent a light bulb that is off by the symbol 0 and a light bulb that is on by the symbol 1, then we can convey a particular pattern by a string of eight 0’s and 1’s. The string 10011101, for example, represents the sequence on-off-off-on-on-on-off-on.

Although a particular pattern of 0’s and 1’s unambiguously conveys the character associated with it, it is useful to interpret this pattern alternatively as an integer in the binary (base-2) number system—a system in which only the two characters 0 and 1 are used to express numbers. In the more familiar decimal (base-10) number system, the digits of an integer reckoned from right to left are the 1’s digit (10 to the zero power), the 10’s digit (10 to the first power), the 100’s digit (10 squared), etc. Similarly, in the binary number system, the bits in an eight-bit integer, again reckoned from right to left, are the 1’s bit (2 to the zero power), the 2’s bit (2 to the first power), the 4’s bit (2 squared), the 8’s bit (2 cubed), the 16’s bit (2 to the fourth power), the 32’s bit (2 to the fifth power), the 64’s bit (2 to the sixth power), and the 128’s bit (2 to the seventh power). Just as the decimal integer 324 means
\[ 3 \times 10^2 + 2 \times 10^1 + 4 \times 10^0 \]
the binary integer 10011101 means
\[ 1 \times 2^7 + 0 \times 2^6 + 0 \times 2^5 + 1 \times 2^4 + 1 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 \]
or, converting to decimal,
\[ 1 \times 128 + 0 \times 64 + 0 \times 32 + 1 \times 16 + 1 \times 8 + 1 \times 4 + 0 \times 2 + 1 \times 1 = 157 \]

---

5The number eight is, of course, arbitrary but conventional. Because computers work internally in the binary (base-2) number system, powers of two—8 = 2³—are especially convenient.

6Actually, the codes will be sent as a stream of bits, each of which is an electrical voltage level that will be either “high” or “lo”, often said to be “on” or “off”. 
The largest three-digit decimal integer is 999; the largest eight-bit binary integer is 11111111, which translates to the decimal integer 255.

An array of eight bits—called a byte—can assume 256 different patterns or values (00000000, 00000001, 00000010, ..., 11111111). Our choice of the byte for internal coding therefore permits us to distinguish 256 codes. Internally, the CPU sees only binary patterns (light bulbs that are on or off; electrical signals that are either high or low; areas on a magnetic tape that are either magnetized or unmagnetized; etc.), and these patterns are conveniently represented by sequences of bits. Externally, binary integers are cumbersome, so various more compact representations are often used. The binary pattern can be interpreted as a decimal integer (as above), but the conversion from binary to decimal is awkward. A more convenient but still compact notation involves grouping the bits in an eight-bit binary integer in the pattern xx-xxx-xxx and using the eight symbols 0, 1, 2, ..., 7 to represent the three-bit binary integers 000, 001, 010, 011, 100, 101, 110, and 111. The integer 10011101, for example, would then have the translation

\[ 10011101 = 10-011-101 = 235 \]

into this octal (base-8) number system. (The first grouping has only two bits and hence can have only the values 0, 1, 2, or 3.) Here, the octal integer 235 is interpreted in decimal as \( 2 \times 8^2 + 3 \times 8^1 + 5 \times 8^0 = 2 \times 64 + 3 \times 8 + 5 \times 1 = 157 \). The largest eight-bit binary integer 11111111 has the representation 377 in octal. This is, of course, the same integer as 255 (decimal).

A still more compact representation of an eight-bit binary integer involves dividing the byte into two four-bit nybbles. Then, with the representation 0000=0, 0001=1, 0010=2, 0011=3, 0100=4, 0101=5, 0110=6, 0111=7, 1000=8, 1001=9, 1010=A, 1011=B, 1100=C, 1101=D, 1110=E, and 1111=F, the binary integer can be represented by two “digits”. For example, the integer 10011101 = 1001-1101 = 9D. This representation expresses the integer in a base-16 or hexadecimal number system. The largest eight-bit binary integer 11111111 has the translation FF into hexadecimal, a value to be compared with 377 in octal and 255 in decimal.

### 1.1.3 The ASCII Character Set

The code transmitted by a particular key on the keyboard is determined by the electrical structure of the keyboard, not by the label on the key. A given key transmits a particular code regardless of the label on the key. Likewise, a code received by a printer identifies, for example, a particular orientation of the printwheel regardless of what character happens to be embossed on the finger at that position. The codes merely identify positions on the keyboard or orientations of the printwheel; no code has any necessary connection with any particular character, and in some contexts associations other than the conventional are adopted.

There are, however, a number of conventional associations of codes with characters. The most commonly used scheme is the American Standard Code for Information Interchange (ASCII, pronounced ass’-key). In this code, characters are associated with eight-bit binary patterns. While the second 128 of the 256 distinguishable patterns [i.e., characters 128–255 (decimal)] have a variety of assignments to characters, the first 128 patterns [i.e., characters 0–127 (decimal)] have the standard assignments enumerated in Table 1.1. The control characters (non-printing characters) all have (decimal) ASCII codes in the range 0–31. Further, the ASCII code for each uppercase letter is 32 less than the code for the corresponding lowercase letter; i.e., turning off the 32-bit in the code for a lowercase letter generates the code for the corresponding uppercase letter. Finally, the ASCII code for a control character is 64 less than the code for the associated uppercase letter; i.e., turning off the 64-bit in the code for an uppercase letter (say C) generates the code for the corresponding control character ([CTRL/C]). Numerical digits occur in ascending order and before the characters in the alphabet; punctuation marks and other symbols (+, -, *, /, @, [, ..., ) are distributed where the previous assignments leave gaps.
CHAPTER 1. PRELIMINARIES

1.1.4 Representation of Data in a Computer

A computer consisting of no more than a keyboard, a CPU with only the above described capabilities, and a printer would, of course, be of little value. Let us expand our computer by adding an internal storage capacity (memory and auxiliary hard disks\(^7\)) consisting of individual cells, each identified by its address, which simply counts the cell’s position from the first cell, and each capable of storing (the code for) a single character. Further, let us endow the CPU with an ability to write codes to and read codes from individual cells in this memory. We understand that a (new) code written into a cell always replaces or overwrites the (previous) contents of that cell, thereby rendering the previous contents no longer retrievable. We declare, however, that reading a code from a cell does not change the contents of the cell.

Typical present-day computers will have a capacity to store an enormous number of bytes—gigabytes, even terrabytes—of information. As we have described it so far, each byte stores an eight-bit pattern of 0’s and 1’s, each pattern being associated with a particular (printing or control) character. The association with characters, however, is not the only possible interpretation of the information stored in one or more bytes of a computer’s memory. Several other interpretations are necessary. Beyond the association of eight-bit patterns with characters (and successions of such patterns with character strings), the CPU might represent integers of various sizes by interpreting

- an eight-bit byte as an unsigned eight-bit integer, assigning its 256 different patterns to the (positive) integers ranging (in decimal) from 0 to 255.
- an eight-bit byte as a signed eight-bit integer, assigning its 256 different patterns to the (negative and positive) integers ranging (in decimal) from −128 to +127. (The range is not symmetric because we must assign one of the patterns to the integer 0.) The highest order bit normally conveys the sign of the value and the remaining seven bits convey the value, though the connection between bit patterns and values—especially negative values—is not always as straightforward as one might naively assume.\(^8\)
- a sixteen-bit combination of two consecutive bytes as an unsigned sixteen-bit integer, with its \(256^2 = 65536\) values assigned to the (positive) integers ranging (in decimal) from 0 to 65536.
- a sixteen-bit combination of two consecutive bytes as a signed sixteen-bit integer, with its 65536 values assigned to the (positive) integers ranging (in decimal) from −32768 to +32767.
- a 32-bit combination of four consecutive bytes as an unsigned 32-bit integer, with its \(65536^2 = 4294967296\) values assigned to the (positive) integers ranging (in decimal) from 0 to 4294967296.
- a 32-bit combination of four consecutive bytes as a signed 32-bit integer, with its 4294967296 values assigned to the (positive) integers ranging (in decimal) from −2147483684 to +2147483683.

Some architectures even use 64-bit unsigned and signed integers to expand the range of available integers even further.

Especially for scientific computations, integers alone will not suffice. Computers provide for storage of numbers with decimal points and exponents by designing the CPU to interpret

- a 32-bit combination of four consecutive bytes as a single-precision floating point number. In the IEEE standard for this format, eight bits (one byte) are assigned to store the proper

\(^7\)For purposes of this discussion we will ignore the very considerable differences between (volatile) memory and (non-volatile) hard disks.

\(^8\)Negative values are frequently stored in what is called two’s-complement form, a discussion of which is beyond the needs or scope of this book. (The two’s complement notation is adopted because it simplifies algorithms that perform arithmetic on signed integers.)
1.1. AN ORIENTATION TO COMPUTERS

Table 1.1: The ASCII character codes. In this table, the first column in each pair lists the decimal code for the character that is identified in the second column of each pair.

<table>
<thead>
<tr>
<th>Code</th>
<th>Character</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>NUL</td>
</tr>
<tr>
<td>16</td>
<td>DLE</td>
</tr>
<tr>
<td>32</td>
<td>SP</td>
</tr>
<tr>
<td>48</td>
<td>0</td>
</tr>
<tr>
<td>64</td>
<td>@</td>
</tr>
<tr>
<td>80</td>
<td>P</td>
</tr>
<tr>
<td>96</td>
<td>i</td>
</tr>
<tr>
<td>112</td>
<td>p</td>
</tr>
<tr>
<td>1</td>
<td>SOH</td>
</tr>
<tr>
<td>17</td>
<td>DC1</td>
</tr>
<tr>
<td>33</td>
<td>!</td>
</tr>
<tr>
<td>50</td>
<td>2</td>
</tr>
<tr>
<td>66</td>
<td>B</td>
</tr>
<tr>
<td>82</td>
<td>R</td>
</tr>
<tr>
<td>98</td>
<td>b</td>
</tr>
<tr>
<td>114</td>
<td>r</td>
</tr>
<tr>
<td>2</td>
<td>STX</td>
</tr>
<tr>
<td>18</td>
<td>DC2</td>
</tr>
<tr>
<td>34</td>
<td>&quot;</td>
</tr>
<tr>
<td>52</td>
<td>4</td>
</tr>
<tr>
<td>68</td>
<td>D</td>
</tr>
<tr>
<td>84</td>
<td>T</td>
</tr>
<tr>
<td>100</td>
<td>d</td>
</tr>
<tr>
<td>116</td>
<td>t</td>
</tr>
<tr>
<td>3</td>
<td>EOX</td>
</tr>
<tr>
<td>19</td>
<td>DC3</td>
</tr>
<tr>
<td>35</td>
<td>#</td>
</tr>
<tr>
<td>53</td>
<td>5</td>
</tr>
<tr>
<td>69</td>
<td>E</td>
</tr>
<tr>
<td>85</td>
<td>U</td>
</tr>
<tr>
<td>101</td>
<td>e</td>
</tr>
<tr>
<td>117</td>
<td>u</td>
</tr>
<tr>
<td>4</td>
<td>EOT</td>
</tr>
<tr>
<td>20</td>
<td>DC4</td>
</tr>
<tr>
<td>36</td>
<td>$</td>
</tr>
<tr>
<td>52</td>
<td>4</td>
</tr>
<tr>
<td>68</td>
<td>D</td>
</tr>
<tr>
<td>84</td>
<td>T</td>
</tr>
<tr>
<td>100</td>
<td>d</td>
</tr>
<tr>
<td>116</td>
<td>t</td>
</tr>
<tr>
<td>5</td>
<td>ENQ</td>
</tr>
<tr>
<td>22</td>
<td>SYN</td>
</tr>
<tr>
<td>38</td>
<td>&amp;</td>
</tr>
<tr>
<td>54</td>
<td>6</td>
</tr>
<tr>
<td>70</td>
<td>F</td>
</tr>
<tr>
<td>86</td>
<td>V</td>
</tr>
<tr>
<td>102</td>
<td>f</td>
</tr>
<tr>
<td>118</td>
<td>v</td>
</tr>
<tr>
<td>7</td>
<td>BEL</td>
</tr>
<tr>
<td>23</td>
<td>ETB</td>
</tr>
<tr>
<td>39</td>
<td>'</td>
</tr>
<tr>
<td>55</td>
<td>7</td>
</tr>
<tr>
<td>71</td>
<td>G</td>
</tr>
<tr>
<td>87</td>
<td>W</td>
</tr>
<tr>
<td>103</td>
<td>g</td>
</tr>
<tr>
<td>119</td>
<td>w</td>
</tr>
<tr>
<td>8</td>
<td>BS</td>
</tr>
<tr>
<td>24</td>
<td>CAN</td>
</tr>
<tr>
<td>40</td>
<td>(</td>
</tr>
<tr>
<td>56</td>
<td>8</td>
</tr>
<tr>
<td>72</td>
<td>H</td>
</tr>
<tr>
<td>88</td>
<td>X</td>
</tr>
<tr>
<td>104</td>
<td>h</td>
</tr>
<tr>
<td>120</td>
<td>x</td>
</tr>
<tr>
<td>9</td>
<td>HT</td>
</tr>
<tr>
<td>25</td>
<td>EM</td>
</tr>
<tr>
<td>41</td>
<td>)</td>
</tr>
<tr>
<td>57</td>
<td>9</td>
</tr>
<tr>
<td>73</td>
<td>I</td>
</tr>
<tr>
<td>89</td>
<td>Y</td>
</tr>
<tr>
<td>105</td>
<td>i</td>
</tr>
<tr>
<td>121</td>
<td>y</td>
</tr>
<tr>
<td>10</td>
<td>LF</td>
</tr>
<tr>
<td>26</td>
<td>SUB</td>
</tr>
<tr>
<td>42</td>
<td>*</td>
</tr>
<tr>
<td>58</td>
<td>:</td>
</tr>
<tr>
<td>74</td>
<td>J</td>
</tr>
<tr>
<td>90</td>
<td>Z</td>
</tr>
<tr>
<td>106</td>
<td>j</td>
</tr>
<tr>
<td>122</td>
<td>z</td>
</tr>
<tr>
<td>11</td>
<td>VT</td>
</tr>
<tr>
<td>27</td>
<td>ESC</td>
</tr>
<tr>
<td>43</td>
<td>+</td>
</tr>
<tr>
<td>59</td>
<td>:</td>
</tr>
<tr>
<td>75</td>
<td>K</td>
</tr>
<tr>
<td>91</td>
<td></td>
</tr>
<tr>
<td>107</td>
<td>k</td>
</tr>
<tr>
<td>123</td>
<td>{</td>
</tr>
<tr>
<td>12</td>
<td>FF</td>
</tr>
<tr>
<td>28</td>
<td>FS</td>
</tr>
<tr>
<td>44</td>
<td>,</td>
</tr>
<tr>
<td>60</td>
<td>&lt;</td>
</tr>
<tr>
<td>76</td>
<td>L</td>
</tr>
<tr>
<td>92</td>
<td>\</td>
</tr>
<tr>
<td>108</td>
<td>l</td>
</tr>
<tr>
<td>124</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>CR</td>
</tr>
<tr>
<td>29</td>
<td>GS</td>
</tr>
<tr>
<td>45</td>
<td>-</td>
</tr>
<tr>
<td>61</td>
<td>=</td>
</tr>
<tr>
<td>77</td>
<td>M</td>
</tr>
<tr>
<td>93</td>
<td>]</td>
</tr>
<tr>
<td>109</td>
<td>m</td>
</tr>
<tr>
<td>125</td>
<td>]</td>
</tr>
<tr>
<td>14</td>
<td>SO</td>
</tr>
<tr>
<td>30</td>
<td>RS</td>
</tr>
<tr>
<td>46</td>
<td>.</td>
</tr>
<tr>
<td>62</td>
<td>&gt;</td>
</tr>
<tr>
<td>78</td>
<td>N</td>
</tr>
<tr>
<td>94</td>
<td>^</td>
</tr>
<tr>
<td>110</td>
<td>n</td>
</tr>
<tr>
<td>126</td>
<td>-</td>
</tr>
<tr>
<td>15</td>
<td>SI</td>
</tr>
<tr>
<td>31</td>
<td>US</td>
</tr>
<tr>
<td>47</td>
<td>/</td>
</tr>
<tr>
<td>63</td>
<td>?</td>
</tr>
<tr>
<td>79</td>
<td>O</td>
</tr>
<tr>
<td>95</td>
<td>_</td>
</tr>
<tr>
<td>111</td>
<td>o</td>
</tr>
<tr>
<td>127</td>
<td>DEL</td>
</tr>
</tbody>
</table>

power or 2 (as an eight-bit signed integer), one bit is assigned to store the sign (0=+), 1=−) of the value, and 23 bits (three bytes minus the sign bit) are assigned to store the digits of the (absolute value of the) value itself. In this format, values ranging (in decimal) from $1.175 \times 10^{-38}$ to $3.403 \times 10^{+38}$ can be represented, though only to a precision of about six decimal digits.\(^9\)

- a 64-bit combination of eight consecutive bytes as a double-precision floating point number.

In the IEEE standard for this format, eleven bits are assigned to store the exponent (as an eleven-bit signed integer, one bit is assigned to store the sign (0=+), 1=−) of the value, and 52 bits are assigned to store the digits of the (absolute value of the) value itself. In this format, values ranging (in decimal) from $2.225 \times 10^{-308}$ to $1.798 \times 10^{+308}$ can be represented, though only to a precision of about fifteen decimal digits.

Must computer architectures conform to these standards. Further, many computers make available one or more extended floating-point formats of their own design.

Clearly, many different data types are in common use. Most importantly, the information stored in a particular byte or aggregate of bytes contains nothing at all to identify its data type. The four bytes of a character string are indistinguishable from the four bytes in a 32-bit unsigned integer and both are indistinguishable from the four bytes in a single-precision floating point number. The bit pattern in those four bytes can be interpreted in any of these ways (and in others as well). It is the programmer’s responsibility to make sure that the program treats stored values in a way appropriate to their data types, usually by referring to memory cells with names that convey the data type. When conversion from one form to another—e.g., character to associated numerical ASCII code—is necessary, the programmer must invoke an appropriate routine to effect the conversion.

\(^9\)Note that the use of an explicit bit to convey the sign of the value means that there are in this format two zeroes. Plus zero is different from minus zero!
1.2 Files and Directories

At (nearly) the most microscopic level, information in a computer is recorded in bytes stored in memory or, more permanently, on a hard drive. At the next level up, aggregations of these bytes into larger units that must be kept together are called files. Each file will have a name. Some of the files containing portions of or referenced in this book, for example, are named \texttt{assemble.tex, laplace.f, laplace.c, trapezoidal.xc, and diffract.ps}. The part of the name before the dot conveys something of what the file contains; when used, the part after the dot—the extension or file type—conveys the type of file.\textsuperscript{10} Some files—called ASCII text files—contain nothing but printing ASCII characters (and perhaps such simple control characters as \texttt{\langle CR\rangle} and \texttt{\langle HT\rangle}) and can be displayed on the screen, printed on a printer, or examined and edited with a text editor. Though some of their bytes can be interpreted as printing characters, other files—called binary files—contain also (perhaps numerous) non-printing characters and cannot be displayed on the screen, examined in (ordinary) text editors, or printed on a printer. Files of this latter type may be special data files created by programs; more often, they are executable files which contain compiled programs, and the bit patterns stored in the file are intended to be interpreted as instructions to the CPU. Whatever the type of file and the nature of the bit patterns it contains, each file is a unit whose component bytes must be kept together as a single entity.

Any computer system will, of course, store a very large number of individual files. To keep these files under some semblance of control, they will commonly be grouped together into aggregates of various sizes, those aggregates will themselves be assembled into higher-level aggregates, those into still higher-level aggregates, \ldots The process is analogous to the aggregating of individual documents into a file folder, of these folders into file drawers, of the drawers into file cabinets, \ldots In the computer world, we need then not only the files themselves but a new type of file that basically lists the contents of the aggregate that it represents. The resulting structure for keeping track of files looks like a tree. At the highest level, the tree has a single file—the root directory—that contains the names of the files it contains (and information about their locations on the disks of the computer). Some of those files may themselves describe (sub)directories, in each of which are listed the names of the files it contains. Some of those files in turn may describe (subsub)directories.\textsuperscript{11} Locating a specific file in the entire structure then requires not only giving the file name but also describing its path—the sequence of directories through which we must pass from the root directory to reach the file. In UNIX, the root directory for the entire storage system is named \texttt{/}; the forward slash is also used to separate directory paths in an extended path. Thus, to specify the location of a file buried several directories down from this universal starting point, we would have to supply an identifier like

\texttt{/usr/people/cook/CCLI/intro/intro.tex}

which indicates that the file \texttt{intro.tex}—the \LaTeX{} source file for this chapter—will be found in the \texttt{intro} directory in the \texttt{CCLI} directory in the \texttt{cook} directory in the \texttt{people} directory in the \texttt{usr} directory in the / (root) directory of the computer system in which it resides.

In the previous paragraph (and in the rest of this book), we use UNIX style file specifications. The corresponding specifications appropriate to the computing system(s) available at your site are described in the \textit{Local Guide}.\textsuperscript{12} That document also explains conventions about (and restrictions imposed on) file names and types, user accounts, and other matters that vary so much from site to site that this book cannot sensibly explain them all.

\textsuperscript{10}The extension \texttt{.tex} conventionally identifies a \TeX{} or \LaTeX{} source file; \texttt{.f} and \texttt{.c} identify files containing source code for FORTRAN and C programs; in this book, \texttt{.xf} and \texttt{.xc} identify executable files generated when FORTRAN and C programs are compiled; and \texttt{.ps}, \texttt{.eps} and \texttt{.pdf} identify PostScript and PDF files.

\textsuperscript{11}Since directories may ultimately be buried many levels deep, we shall suppress the multitude of sub's that might appear, understanding that the simpler word `directory' will refer to a directory without regard to its position in the overall hierarchy.

\textsuperscript{12}In Windows, for example, the backslash character \texttt{\textbackslash} is used to separate directories in a path.
1.3 Operating Systems

Underneath it all, everything that a computer does is controlled by its operating system, which makes available a variety of standard commands to instruct the computer to carry out common tasks. In some cases, the user invokes a command by typing its name (and any necessary arguments) in a text-entry window or command-line interface (CLI). In other cases, the user clicks a mouse button on an icon or drags an icon to a new location on the desktop in a graphical user interface (GUI). However a command is conveyed to the operating system, it at base simply invokes a program that carries out the selected task and then returns control to the operating system for the next command. At the very minimum, the operating system must make available commands for

- logging in and logging out, paying attention on multi-user systems to user authorization (normally controlled through usernames and passwords).
- setting and changing the default directory, which is the directory to which file names refer when no path is specified.
- copying a file to another directory or deleting it altogether from its current directory.
- establishing various levels of file protection file by file and changing those specifications.
- creating ASCII files through the use of a text editor.
- customizing the user’s environment through the creation of environment variables, aliases, and other shorthands.
- retrieving and editing a previously executed command before it is submitted again for execution.
- displaying a file on the screen.
- printing a file to a printer.
- copying a selected portion of the screen to a file.
- converting files from one format to another.

The details of the ways in which these several capabilities are invoked and conventions about assigning user names, passwords, and default directories vary considerably among operating systems and are, even with the same operating system, site-specific. The Local Guide for your site describes those details.

1.4 Glossary, Conventions, and Understandings

In this section, we enumerate and define a number of terms to be used throughout this book, and we make a variety of observations that otherwise would have to be repeated several times.

- Typographically, we use the typewriter font for all program listings and for command lines displayed in the text. We also use this font for command and function names embedded in the text itself without enclosing these names in quotation marks (unless the absence of quotation marks creates ambiguity or confusion).
- In describing mouse operations, we use ML, MM, and MR for the left, middle, and right mouse buttons, respectively.\footnote{If you have invoked a feature of your operating system that permits reversing the conventional association of mouse buttons with actions, then you will have to read our MR to mean your ML, etc.} The Local Guide explains how to translate these symbols if your mouse has fewer than three buttons.
• As a shorthand, we use the phrase ‘Select ...’ for the operations of moving the cursor over the indicated item (which may involve pulling down a menu) and then clicking ML.

• We use *italic* type for window names, **small caps** to identify menus, and single quotation marks to enclose the names of buttons or menu items. Thus, for example, in a tutorial segment, we might instruct you to “Select ‘Print’ from the FILE menu in the WinEdit window”.

• The lines dividing statements from commands from instructions are difficult to draw. In this book, we strive to refer to a complete instruction in some programming language as a *statement* and to reserve the word *command* for the keyword that introduces a statement. For example, we would speak of the command *integrate* but refer to the construction

\[
\text{integrate( sin(k*x), x, 0, Pi/2 )}
\]

as a statement. Even this distinction is difficult to draw, however, because statements can be nested to produce compound statements that could, with justification, themselves be referred to as statements.

• The lines dividing functions from procedures from subroutines are also difficult to draw. Indeed, some computer languages regard these terms as synonymous. When a distinction is made, a function is a construction which, when executed, accepts arguments as input but returns a value to the variable(s) to which the function is assigned; the function SQRT, for example, would be invoked with a statement like

\[
R = \text{SQRT( X^2 + Y^2 )}
\]

Procedures and subroutines, however, (normally) have only arguments, some of which will supply input and others of which name the variables into which returned values will be placed; a procedure—call it SQRTPRO—to return in its second argument the square root of its first argument would be invoked with a statement like

\[
\text{SQRTPRO( X^2 + Y^2, R )}
\]

having no variable or equal sign at its beginning.\(^{14}\)

• Most of the statements presented in this book can be submitted as they stand and executed by the program in whose command language the statement is written. Occasionally, we illustrate the general format of a statement without being sufficiently explicit to render the statement executable. Statements in the former category will be preceded by the appropriate prompt; statements in the latter category will be presented without a prompt. As a general rule, statements preceded with a prompt can—and should—be executed as you work your way through the material. Statements without a prompt should not, and most often could not, be executed.

• Especially in constructing statements for computer algebra systems and presenting their output, we will not always present the output in *exactly* the form or with *exactly* the appearance it will actually have. In particular, we will frequently use unsubscripted variables, e.g. \(x_1\) or \(x_f\), in the statement to be executed but render these variables as subscripted, e.g., \(x_1\) or \(x_f\), in the displayed output.

• At various points in this book, we define, use, and/or refer to a variety of files specific to the text. All such files are stored on your local computer system and are available for your use. The head of the directory tree in which those files are stored is identified explicitly in the

---

\(^{14}\)In some languages, procedures can optionally be written \(\text{err = SQRTPRO( X^2+y^2, R )}\), in which case the procedure returns a value to the variable \(\text{err}\) to convey that the procedure encountered a problem in its execution. Testing \(\text{err}\) after the procedure is invoked can then be used to trap errors and alert you to possible incorrect output.
1.5 Assumed Background

Mathematically, we assume in this book that you understand the notions of derivatives, integrals, and ordinary differential equations and that you have some acquaintance with linear algebra (matrix operations, eigenvalues and eigenvectors, ...). Physically, we suppose that you have taken a couple of introductory, calculus-based courses in physics and are continuing with intermediate courses in physics. This book makes contact with many intermediate-level physics courses, but it is not focused on any particular one of those courses. It draws on topics covered in several such courses whenever appropriate.
In addition, we include here a brief discussion of two mathematical topics that are necessary for some of what follows but that may well not have been treated in any of the courses viewed as prerequisite for the study of this book.

1.5.1 The Gamma Function

The factorial function \( n! \), which is defined when \( n \) is an integer as the product of all integers from \( n \) down to 1, i.e., by
\[
n! = n \times (n-1) \times (n-2) \times (n-3) \times \cdots \times 2 \times 1 \tag{1.1}
\]
is usually familiar. The double factorial function \( n!! \), which is defined (again when \( n \) is an integer) as the product of every other integer, i.e., by
\[
n!! = n \times (n-2) \times (n-4) \times (n-6) \times \cdots \times \begin{cases} 4 \times 2 & n \text{ even} \\ 3 \times 1 & n \text{ odd} \end{cases} \tag{1.2}
\]
is less familiar but occurs frequently as well. Whether \( n \) is even or odd, the double factorial can be recast in terms of single factorials. If, for example, \( n \) is even, say 10, we can recast its double factorial in the form
\[
10!! = 10 \times 8 \times 6 \times 4 = 2^5 \times 5 \times 4 \times 3 \times 2 = 2^5 5! \Rightarrow (2n)!! = 2^n n!
\tag{1.3}
\]
If \( n \) is odd, say 9, recasting its double factorial takes mildly more work, but is illustrated in the chain
\[
9!! = 9 \times 7 \times 5 \times 3 \times 1 = \frac{9 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{8 \times 6 \times 4 \times 2} = \frac{9!}{8!!} = \frac{9!}{2^4 4!}
\Rightarrow (2n+1)!! = \frac{(2n+1)!}{2^n n!} \tag{1.4}
\]

We might wonder whether it is possible to define a function of a continuous variable that will coincide with the factorial function when its argument is an integer. Consider the function \( \Gamma(\nu) \) defined by the integral
\[
\Gamma(\nu) = \int_0^{\infty} t^{\nu-1} e^{-t} \, dt \tag{1.5}
\]
We quickly conclude that
\[
\Gamma(1) = \int_0^{\infty} e^{-t} \, dt = 1 \tag{1.6}
\]
With a little more effort, we can evaluate \( \Gamma(1/2) \). We begin by writing the definition of \( \Gamma(1/2) \) and introducing the new variable \( x^2 = t \) in the definition, finding that
\[
\Gamma \left( \frac{1}{2} \right) = \int_0^{\infty} \frac{e^{-t}}{\sqrt{t}} \, dt = \int_0^{\infty} \frac{e^{-x^2}}{x} \, dx = \int_0^{\infty} e^{-x^2} \, dx \tag{1.7}
\]
Then, we examine the square of the quantity of interest, change to polar coordinates, and find that
\[
\left[ \Gamma \left( \frac{1}{2} \right) \right]^2 = \left( \int_0^{\infty} e^{-x^2} \, dx \right)^2 = \left( \int_0^{\infty} e^{-x^2} \, dx \right) \left( \int_0^{\infty} e^{-y^2} \, dy \right) \tag{1.8}
\]
\[
= \int_0^{\infty} \int_0^{\infty} e^{-(x^2+y^2)} \, dx \, dy = \int_0^{\infty} \int_0^{2\pi} e^{-r^2} \, r \, dr \, d\phi = \pi \tag{1.9}
\]
and we conclude that
\[ \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi} \] (1.10)

Evaluation of the Gamma function at most other arguments must be done numerically.

The Gamma function, however, has a particularly interesting property that we can deduce if we apply integration by parts to the definition. Provided \( \nu > 1 \), we find that
\[
\Gamma(\nu) = -\int_0^{\infty} t^{\nu-1} e^{-t} \bigg|_0^{\infty} + (\nu - 1) \int_0^{\infty} t^{\nu-2} e^{-t} dt = (\nu - 1) \Gamma(\nu - 1) \] (1.11)

Applying this recursion relationship when the argument of the Gamma function is an integer, we find, for example, that
\[
\Gamma(5) = 4 \Gamma(4) = 4 \times 3 \Gamma(3) = 4 \times 3 \times 2 \Gamma(2) = 4 \times 3 \times 2 \times 1 \Gamma(1) = 4 \times 3 \times 2 \times 1 = 4! \] (1.12)

More generally, a similar argument leads to the conclusion that
\[
\Gamma(n + 1) = n! \] (1.13)

and we have indeed succeeded in finding a function that is the natural extension of the factorial function to non-integral arguments. Indeed, one often sees the notation \( \nu! \) as an alternative to the notation \( \Gamma(\nu + 1) \)—and the latter in fact provides a formal definition of the former.\(^{15}\) Note that, since we know quite explicitly that \( \Gamma(1) = 1 \), this connection between the Gamma and factorial functions supports what is sometimes an assertion of convenience, namely that \( 0! = 1 \).

### 1.5.2 The Laplace Transform

One tool used behind the scenes by symbolic solvers of ordinary differential equations is called the Laplace transform, which we describe here to avoid duplicating the discussion at several places in subsequent chapters. While we are not likely to make much use of the Laplace transform directly, knowing its properties may sometimes be valuable as we try to guide a symbolic manipulator that uses the technique. Defined for a function \( f(t) \) by the integral
\[
\mathcal{L}\left(f(t)\right) = \tilde{f}(s) = \int_0^{\infty} e^{-st} f(t) dt \] (1.14)

this transform has several important properties:

- The Laplace transform of a linear combination of functions is that same linear combination of the Laplace transforms of the separate functions,
\[
\mathcal{L}\left(af(t) + bg(t)\right) = \int_0^{\infty} e^{-st} \left(af(t) + bg(t)\right) dt \\
  = a \int_0^{\infty} e^{-st} f(t) dt + b \int_0^{\infty} e^{-st} g(t) dt \\
  = a\mathcal{L}\left(f(t)\right) + b\mathcal{L}\left(g(t)\right) \] (1.15)

  i.e., in more technical terminology, \( \mathcal{L} \) is a linear operator (because integration itself is a linear operation).

\(^{15}\)The requirement at Eq. (1.11) that \( \nu > 1 \) limits the range of \( \nu \) for which the integral is acceptable as a definition of the Gamma function. Outside that range, we simply take the recursion relationship itself to define the function, so the recursion relationship is always valid while the integral converges only for \( \nu > 1 \).
Table 1.2: A short table of Laplace transforms.

<table>
<thead>
<tr>
<th>$f(t)$</th>
<th>$\tilde{f}(s)$</th>
<th>$f(t)$</th>
<th>$\tilde{f}(s)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t^n$</td>
<td>$\frac{n!}{s^{n+1}}$</td>
<td>$\sin \omega t$</td>
<td>$\frac{\omega}{s^2 + \omega^2}$</td>
</tr>
<tr>
<td>$\cos \omega t$</td>
<td>$\frac{s}{s^2 + \omega^2}$</td>
<td>$e^{at}$</td>
<td>$\frac{1}{s - a}$</td>
</tr>
<tr>
<td>$\frac{dx}{dt}(t)$</td>
<td>$s \tilde{x}(s) - x(0)$</td>
<td>$\frac{d^2x}{dt^2}(t)$</td>
<td>$s^2 \tilde{x}(s) - s x(0) - \frac{dx}{dt}(0)$</td>
</tr>
</tbody>
</table>

- The Laplace transform of the first derivative of a function $f(t)$ is simply related to the Laplace transform of $f(t)$. We need merely integrate the formal expression for the transform of the derivative by parts to find that

$$\frac{df}{dt}(s) = \int_0^\infty e^{-st} \frac{df(t)}{dt} dt = e^{-st} f(t) \bigg|_0^\infty + s \int_0^\infty e^{-st} f(t) dt = s \tilde{f}(s) - f(0) \quad (1.16)$$

- The Laplace transform of a higher-order derivative is also simply related to the Laplace transform of the original function. We merely apply the identity in Eq. (1.16) repeatedly. The Laplace transform of a second derivative, for example, has the evaluation

$$\frac{d^2f}{dt^2}(s) = s \frac{df}{dt}(s) - \frac{df}{dt}(0) = s \left(s \tilde{f}(s) - f(0)\right) - \frac{df}{dt}(0) = s^2 \tilde{f}(s) - sf(0) - \frac{df}{dt}(0) \quad (1.17)$$

As we shall see particularly in the chapter on ordinary differential equations, these last two properties, which convert differential expressions involving $f(t)$ into algebraic expressions involving $\tilde{f}(s)$, can be extended to convert some types of differential equations into algebraic equations. As a consequence, we anticipate that the Laplace transform may well play an important role in some approaches to solving ordinary differential equations.

Provided we can actually do the integral in Eq. (1.14), we can, of course, supplement these general properties by explicit evaluation of any number of Laplace transforms. Each entry in Table 1.2—a very short table of Laplace transforms—was obtained by explicit evaluation of the defining integral for the corresponding function.

1.6 Licensing Issues

Much of the software on every device in computational facilities around the world is proprietary and subject to the provisions both of the applicable copyright laws and of license agreements between the local institution and the vendors of the software. Usually—but not always, the licenses acquired by a given institution will permit simultaneous use on all of the devices in a laboratory at that institution. Almost certainly, the licenses limit use to projects and activities at that institution and prohibit copying of the software, except for purposes of system maintenance and backup. All users of all devices must be constantly mindful of the proprietary nature of much of the available software and must abide by the restrictions imposed by the copyright laws and by the license agreements. Those restrictions for each software package available at your site are described in the Local Guide.
Chapter 4

Introduction to OCTAVE

Note: All OCTAVE program (*.m) and data (*.dat) files referred to in this chapter are available in the directory $HEAD/octave, where (as defined in the Local Guide) $HEAD must be replaced by the appropriate path for your site. At some sites, this directory or some other directory containing these files may also have been placed in OCTAVE’s default search path. (See Section 4.18.4.) If so, the files can be found by OCTAVE without explicit specification of a path. Otherwise, you will have to use the full path to copy them into your default directory to access them.

OCTAVE\(^1\) is an interactive software package (similar in function and syntax to the commercially available program MATLAB) whose primary—though certainly not only—capabilities are to manipulate arrays, which may be scalars (\(1 \times 1\) arrays), one-dimensional vectors (\(1 \times n\) or \(n \times 1\) arrays), or two-dimensional matrices (\(m \times n\) arrays), and to produce graphical visualizations. OCTAVE works in conjunction with a graphics toolkit—three different toolkits are available—to plot figures. A typical interaction with the program will involve (1) creating one or more arrays in OCTAVE’s workspace (either by computing them directly in OCTAVE or by importing them into OCTAVE from a file created by another program), (2) processing them in some way to produce other arrays, and (3) displaying the end results in an appropriate graphical form. In anticipation of such uses in later chapters, this introduction describes the elementary commands for creating, processing, and displaying arrays. Further details can be found in the online help messages accessible from within OCTAVE itself (see Section 4.8) and in assorted OCTAVE documentation (see Section 4.19). We shall refer to this documentation collectively as the OCTAVE manuals.

Two user interfaces to OCTAVE are available. The command-line interface (CLI) is the simpler of the two and runs directly in a shell window of the operating system. When available\(^2\) and invoked, the graphical user interface (GUI) comes up in its own window. Commands to OCTAVE are typed in the same way in both interfaces, but the GUI provides added features that facilitate constructing commands, keeping track of variables used, creating command files, accessing documentation, \ldots. In either interface, when graphical displays are requested, OCTAVE will open a separate Plot window, which will display the graph and offer a few options for adjusting the display, though a variety of OCTAVE commands can also be invoked to adjust the display.

\(^1\)Octave, created primarily by John W. Eaton of the University of Wisconsin, is named after one of the author’s former professors, Octave Levenspiel and is a freely available program at \url{www.octave.org}. One is free to redistribute and/or modify the programs under the terms of the GNU General Public Library (GPL) as published by the Free Software Foundation. OCTAVE encourages users to contribute useful functions to expand OCTAVE’s library.

\(^2\)The GUI is not available on all platforms.
4.1 Beginning an OCTAVE Session

Detailed instructions for initiating a session with OCTAVE will be found in the *Local Guide*. Usually, OCTAVE will be started either (1) by typing a command at the prompt from the operating system,\(^3\) (2) by double-clicking the left mouse button on an appropriate icon on the desktop, or (3) by selecting OCTAVE from a menu. Throughout the course of this chapter—either set by default or explicitly if necessary—we assume that OCTAVE is run with the modifier *--traditional*.\(^4\) Presently, the OCTAVE prompt `>>` will appear, either in the **Shell** window of your operating system (CLI) or in a newly created **OCTAVE Command Window** (GUI). *Ab initio*, we must be aware that

1. Internally, OCTAVE is case sensitive. A and a do not have the same meaning.

2. Typing the command `quit` or `exit` at the OCTAVE prompt ends the session.

3. `(Control-C)` instructs OCTAVE to abort its present activity. The activity is not aborted, however, until the statement that was active finishes. Thus, it may take a while before the OCTAVE prompt returns.

4. Every command line in OCTAVE must end either with a `{RETURN}`, which triggers the execution of the statement(s) on the line and the display of a suitable response, or a semicolon `;' followed by a `{RETURN}`, which triggers the execution of the statement(s) on the line but suppresses the display of the response. If a statement is too long to fit comfortably on one line, the symbol `...` at the end of one line can be used to tell OCTAVE that the statement is continued on the next line.

5. Commas or semicolons can be used to separate independent statements on a single line. Using commas will display the output of the preceding command while using semicolons will suppress the output of the preceding command.

6. In any command line, the pound (number) sign `#` or, alternatively, the percent sign `%` can be used to introduce a comment. The pound or percent sign and all following characters in the line will be ignored. Several lines can be commented by enclosing them in either `#{...#}` or `%{...%}`. We shall use only `#`.

7. The OCTAVE command `system` allows temporary escape from OCTAVE to the operating system. Thus, for example, the statement `!system("ls *.ps")` (UNIX) or `!system("dir *.ps")` (Windows) issued to OCTAVE will be executed not by OCTAVE but rather by the underlying operating system. Control is then returned to OCTAVE. If you are working in the GUI or can launch more than one **Shell** window, this feature is, of course, unnecessary—since we can easily use resources of the desktop to execute other programs without exiting from OCTAVE.

8. Within OCTAVE, we can retrieve previous statements with the up-arrow key and edit individual statements with the backspace key and the left and right arrow keys. Additionally, with some interfaces on some platforms, we can copy all or portions of already executed statements to the current input point in the command window by highlighting the desired text, clicking the right mouse button, selecting ‘copy’ from the resulting pop-up menu, clicking the right mouse button again, and selecting ‘paste’ from the resulting pop-up menu. Further, pressing {RETURN} will execute the entire command line, regardless of where the cursor happens to be positioned within the line.

\(^3\)If your system manager has made this command available, this route will allow the specification of additional modifiers that are enumerated in the OCTAVE help documents. In this command to the operating system, case may be important.

\(^4\)See the *Local Guide* to learn if your system manager has implemented a different default behavior. In particular, OCTAVE’s prompt for commands may be modified at some sites.
4.2 Basic Entities in OCTAVE

4.2.1 Data Types

While individual items of data can be aggregated into quite complicated structures and OCTAVE admits declaration of several data types or classes, by default only two basic data types are likely to be used with any frequency. Whether they have integral or decimal values, numbers are all by default stored as double-precision, floating point values, so each number requires eight bytes and a $2 \times 3$ array of such numbers, for example, will have six elements and require 48 bytes for its storage. Values will be stored to approximately sixteen decimal digits and can have magnitudes ranging from about $10^{-308}$ to about $10^{308}$. OCTAVE refers to this data class as double. For character strings (class char), on the other hand, OCTAVE allocates two bytes per character, so an eight-character string, for example, will occupy 16 bytes of storage.

We need take no particular action to define the data type of any variable. OCTAVE determines an appropriate data type automatically and dynamically. Thus, a particular variable may have one type at one point in a session and a different type at another point in the same session.

With one exception, the OCTAVE commands num2str, int2str, toascii, and str2num, can be used to convert a numeric value to a string of characters, an integer value to a string, a string to a vector of ASCII code equivalents, and a string to a numeric value, respectively. In addition, the commands fix, ceil, floor, and round can be used to convert a floating point value into an integer value, where fix rounds towards zero, ceil and floor round toward $+\infty$ or $-\infty$, respectively, and round rounds up or down to the nearer integer. During some conversions, e.g., the conversion of the floating point value 3.14 with the expression int2str(3.14), which will yield the string '3', information will be lost in the rounding and cannot be retrieved by invoking the reverse command str2num. Converting a string to a number can be done by using the str2num command if the characters within the string are all digits, or with the function toascii, which converts the vector of characters to their ASCII representations and puts those integers into a vector, one element per character.

While OCTAVE by default assigns a data type to each entered value, the user can force use of alternative data types through the invocation of one or another commands. Specifically,

- double(a) converts a to double precision floating point
- single(a) converts a to single precision floating point
- int8(a) converts a to an 8-bit signed integer
- uint8(a) converts a to an 8-bit unsigned integer

and similarly for the commands int16, uint16, int32, uint32, int64, and uint64.

4.2.2 Variable Names

The simplest variables in OCTAVE represent either scalars (single values) or arrays (several values, all of the same data type, referred to by a single name). Variable names must start with a letter and can have any number of characters, though names longer than about 30 characters are not recommended. Upper- and lower-case letters, numbers, and the underscore character are legal. However, variables that begin and end with two underscores are understood to be reserved for internal use by OCTAVE. One should not use variables of this type when writing code unless trying to access OCTAVE’s internal variables. Variable names are sensitive to case. Because the data type of a variable is assigned dynamically, a particular variable may have one type at one point in a session and a different type at another point in the same session.

---

5A string value (e.g., 'cook') that does not represent a valid number cannot be converted to a number.
4.2.3 Assignment of Values to Variables

In OCTAVE, the equal sign `=` plays the role of the assignment operator so, for example, the statement

```
alpha = 3.79
```

creates a variable of class `double` named `alpha` and assigns the value 3.79 to that variable, i.e., stores in an eight-byte memory location named `alpha` the (double precision) floating point binary representation of the value 3.79. Similarly, the statement `name = 'David'` or `name = "David"` creates a variable of class `char` and assigns the value `David` to that variable, i.e., stores the codes for each of the five characters in a five element array, each element of which occupies two bytes. Even though the distinction between an integer, say 3, and a floating point number, say 3.0, that happens to have an integral value has no bearing on the storage format of that value, placing explicit decimal points in all floating point values to highlight the distinction is nonetheless recommended. Note also (1) that string constants are enclosed in single or double quotation marks and (2) that the values returned by an OCTAVE statement will be automatically displayed unless the statement is terminated with a semicolon.

Beyond setting variables equal to specific values, a variable can be set equal to an expression with the understanding that OCTAVE will evaluate the expression to determine the value to be assigned to the variable. To achieve that end, of course, all variables used in the construction of an expression must have previously been assigned explicit numerical values. Thus, for example, the statement

```
x = alpha + beta
```

will add the values currently stored in the variables `alpha` and `beta` and store the result in the variable `x`, creating `x` if the variable does not already exist and overwriting whatever is currently stored in `x` if the variable already exists.

4.2.4 Commands

In OCTAVE the variables identify storage areas in memory and provide ways to refer symbolically to the values stored in those areas. Actual processing of values is effected by one or another of OCTAVE’s `commands`, the behavior of which is almost always influenced by the value or values of one or more arguments and/or `properties`. Statements invoking a particular command will usually be expressed in the form

```
command( argument, argument, 'PropertyName', PropertyValue, ...)
```

where the arguments are separated by commas. The statement begins with the command name, which will be followed by one or more arguments, all enclosed in a single set of parentheses. Some of these arguments are “free” and others are specified by using a named property. Most commands, however, return values to the program that invokes them and, if we wish to preserve the returned values for later use, we should invoke the command with a statement in the general format

```
q = command( argument, argument, 'PropertyName', PropertyValue, ...)
```

in which the value returned by the command is assigned to the user-specified variable `q`—though OCTAVE will assume the variable `ans` if no variable is specified. The order of the “free” arguments is mandatory, since the position of each argument identifies its role. Arguments specified by properties, however, can be presented in any order, since the property name identifies the role of the immediately following value. Most properties have appropriate default values, so their explicit stipulation is necessary only if the default value is unacceptable. Property names, which are always strings, those property values that are strings, and any arguments that are strings are always enclosed in `single` quotation marks; numeric arguments and numeric property values are not quoted.
In the previous paragraph, we described the construction of statements in which the arguments are presented within parentheses and each statement looks very much like the evaluation of a function in conventional mathematical notation. Some commands, however, can accept arguments either with or without the parentheses. Thus, for example, the statement `hold('on')`, which turns off automatic erasing of the current graph when a subsequent plot is created in the same window, can alternatively be submitted to OCTAVE in the form `hold on`. Full information about this feature, which OCTAVE refers to as command/function duality can be found in the OCTAVE manuals.

### 4.3 A Sampling of OCTAVE Capabilities

In this section, we present several examples illustrating various capabilities of OCTAVE and introducing some of the most frequently used commands.

#### 4.3.1 Creating and Examining Arrays

Arrays can be created within OCTAVE or—as seen in Section 4.7—read in from files. Individual elements of arrays can be accessed by specifying the row and column, as in `array_name(row,column)`, where the first index identifies the row, the second index identifies the column, the indices are enclosed in parentheses, and they begin at one (repeat: indices begin at 1). Alternatively, an element in an array can be identified by a single number, e.g., as in `array_name(element number)`. For this form, the elements in the array are numbered sequentially starting with one in the upper left of the array and then continuing down each column before moving right to the top of the next column. The syntax for creating arrays, assigning values to elements in arrays, and referring to arrays using subscripts is illustrated in the following paragraphs.

The simplest assignments assign single values to variables. We might, for example, create and display simple string and numeric variables with statements like

```octave
>> name = 'David'
name = David

>> val = 3.1415926535
val = 3.1416
```

Assign string value to `name` and print it to the screen. The actual variable will be a 1 × 5 string array.

```octave
>> val = 3.1415926535
val = 3.1416
```

Assign numeric value to `val` and print it to the screen. The actual variable will be a 1 × 1 matrix of class double. All digits are stored internally using double precision. By default, for display, OCTAVE adopts its default format to display five significant digits.

```octave
>> cmp = 3.0+5.0*i
cmp = 3+5i
```

Assign a complex value to the variable `cmp`, using the special symbol `i` for \( \sqrt{-1} \). OCTAVE also recognizes the special symbols \( I, j, \) and \( J \) for \( \sqrt{-1} \), though all will be displayed as `i` in output.

To create vectors and arrays containing more than one numeric element, we must use slightly more complicated statements like

---

6 We here limit our discussion to arrays of one and two dimensions. See the OCTAVE manuals for ways to create arrays of three or more dimensions.

7 OCTAVE statements are shown on the left; comments describing the statements are shown on the right. Further, to save space, we will routinely compact OCTAVE’s output by omitting blank lines and extra spaces.
Create a (row) vector \( b \), separating elements in a row with spaces, and print it to the screen. Commas may also be used rather than spaces to separate elements of a vector. That is \( b = [1, 0, 3, 7, 10] \) will produce the same vector.

Assign values to a \( 3 \times 3 \) matrix, using a terminating semicolon to suppress output. Then print it to the screen by asserting the variable name. Note the second use of the semicolon to separate rows in an array.

Create a \( 100 \times 100 \) array of zeros, again terminating the statement with a semicolon to suppress a lot of output (indices range from 1 to 100).

Create a \( 3 \times 3 \) matrix (nine elements total), each element being a random number uniformly distributed between 0 and 1. Different versions of OCTAVE may display different numbers of digits by default. See the OCTAVE manuals for ways to customize the behavior of \texttt{rand}.

Other functions that can be used to create arrays include \texttt{eye(n)} (to make an \( n \times n \) unit matrix) and \texttt{magic(n)} (to make an \( n \times n \) magic square). Other famous matrices are enumerated in the OCTAVE manuals.

To remind ourselves of the variables OCTAVE knows at any particular moment and to find out something about those variables, we would use statements like

```
>> who
*** local user variables:
   a   b   cmp   data   name   val   z
```

Ask for list of known variables. Different versions of OCTAVE may display different explanatory messages.

```
>> whos
Variables in the current scope:

   Attr Name   Size   Bytes    Class
     ==== ==== ===== ======
     a   3x3   72 double
     b   1x5   40 double
     c   cmp   1x1   16 double
     data 100x100 80000 double
     name 1x5   5 char
     val   1x1    8 double
     z   3x3   72 double

Total is 10030 elements using 80213 bytes
```

Then, should we wish to delete some of these variables to free memory for other purposes, we could use a statement like

```
>> clear b cmp data name z
```
(or `clear('b', 'cmp', 'data', 'name', 'z')`). Either one of the statements `clear` or `clear *` would remove all known variables.

Once an array has been created, we can access individual elements in the array with statements like

```plaintext
>> a(1,2)  \hspace{1cm} \text{Print the value in the first row, second column of a. OCTAVE uses the variable ans if no "output" variable is specified.}
an = 2
```

Alternatively, exploiting the capacity to address elements of an array with a single index, we could use a statement like

```plaintext
>> a(4)  \hspace{1cm} \text{Display the fourth element of a. Note that this element is the same as the element printed by the previous statement, but it is being addressed by element number rather than by its column and row indices.}
an = 2
```

```plaintext
>> a(4) = 10  \hspace{1cm} \text{Assign the value 10 to the fourth element (first row, second column) of a. The edited array is printed automatically.}
a =
  1  10  3
  4  5  6
  7  8  9
```

```plaintext
>> a(2,:)  \hspace{1cm} \text{Print the second row of a. The colon by itself functions as a wild card and stands for all values of the corresponding index.}
an =
  4  5  6
```

```plaintext
>> c = a(1:2,1:2)  \hspace{1cm} \text{Extract a $2 \times 2$ matrix made from the four elements in the upper left hand corner of a and print it. The colon between two values tells OCTAVE to use a range of values running in integer steps from the first to the last, inclusive.}
c =
  1  10
  4  5
```

Before leaving issues related to creation and display of arrays, we note one more command: the command `format` allows adjustment of the way in which values are displayed. We might adjust that format with statements like

```plaintext
>> format long  \hspace{1cm} \text{Specify the long format and redisplay the value by asserting its name. This time 15 digits will be displayed. Here we use format as a command. We expose OCTAVE’s command/function duality by noting that this statement might also have been written `format('long')`.}
>> val
val = 3.1415926535000
```

```plaintext
>> format  \hspace{1cm} \text{Return to default format.}
```

Other formats include

- `short` which shortens the output to three significant figures,
- `short e/short E`, which forces presentation of numbers in short format but uses the scientific format (3.1415926535 becomes 3.14e+00 or 3.14E+00),
- `long e/long E`, which forces presentation of numbers in long format but uses scientific notation,

Regardless of the display, internally OCTAVE uses full double precision for all calculations and storage.
4.3.2 Simple Mathematical Manipulations

The ordinary mathematical operations of combining of variables and constants are achieved with the usual operators: + for addition, - for subtraction, * for multiplication, / for division, and ^ for exponentiation. Within OCTAVE, these operators are given the usual priorities so that, for example, in the expression a*b^2 + c/d, b is first squared, then a is multiplied by b^2 and c is divided by d, and finally c/d is added to a*b^2. Parentheses can be used to effect an order of operations different from that dictated by the default priorities.

As we shall see in more detail in the next section, OCTAVE understands the operators *, ^, and / applied to matrices to effect matrix multiplication, matrix exponentiation, and matrix division (i.e., multiplication by the inverse matrix). If—as we often do—we want element-by-element multiplication, exponentiation, and division of matrices, we must use OCTAVE’s operators .*, ./, and .^.

Further, OCTAVE is aware of the many functions that are important to scientific computation, including the absolute value abs; the square root sqrt; the trigonometric functions sin, cos, and tan; the inverse trigonometric functions asin, acos, and atan; the base-10 and natural logarithms, log10 and log; the exponential exp; the hyperbolic functions cosh, sinh, and tanh; and the Bessel functions of real—though not necessarily integer—order besselj, besseli, and bessely. Functions may have one or more arguments (see the OCTAVE manuals). Those that have a single argument are invoked with a statement of the form

```
>> q = function_name(x)
```

after the execution of which the variable q will contain the value of the corresponding function at the argument contained in the variable x.\(^8\) Note particularly that the arguments given to the trigonometric functions must be in radians, and the values returned by the inverse trigonometric functions will be in radians. Note also that the function atan2

```
>> q = atan2( y, x )
```

has two arguments and returns in q the value atan(y/x) of the corresponding angle in the proper quadrant as determined by the separate signs of x and y.

Continuing from the point reached at the end of Section 4.3.1, the following record of an OCTAVE session demonstrates a few mathematical manipulations:

```
>> eq1 = 8.4*sin(30.0*pi/180.0)*sqrt(3.0)
    eq1 =  7.2746

>> fix(eq1)
    ans =  7

>> c
    c =  1     10
        4     5

>> d = [ 3 4; 7 4 ]
    d =  3     4
        7     4

>> e = c + d
    e =  4     14
        11     9
```

---

\(^8\)For a properly defined function, x can also be an array. The variable q will then also be an array whose elements are obtained by element-by-element application of the function to the entries in x.
4.3. A SAMPLING OF OCTAVE CAPABILITIES

>> e = c .* d
e = 3 40
  28 20

Multiply c and d element by element (using the operator \( \cdot \)).

>> f = c ./ d
f = 0.33333 2.50000
  0.57143 1.25000

Divide c by d element by element. Note that floating division has been done even though the elements in the matrix had no explicit decimal points. Also note that the evaluation of each element has at least five significant digits and all values have the same precision.

4.3.3 Matrix Manipulations

OCTAVE adheres to the usual convention in which the first index in the symbol \( a(i,j) \) for an element of a matrix is regarded as the row index and the second index as the column index. For matrix multiplication, which is defined by

\[
e = dc \implies e_{ij} = \sum_k d_{ik} c_{kj}
\]  

the summation for OCTAVE involves migrating across the \( i \)-th row (i.e., from column to column in the \( i \)-th row) of the first matrix as we migrate down the \( j \)-th column (i.e., from row to row in the \( j \)-th column) of the second matrix. That process is specified by the operator \(*\) and, for the matrices \( d \) and \( c \) defined above, yields the result

>> e = d * c
e = 19 50
  23 90

Find the matrix product \( dc \).

>> e = c * d
e = 73 44
  47 36

Find the matrix product \( cd \), and note that \( dc \neq cd \).

OCTAVE, of course, knows about all of the common entities that we might wish to determine from a single matrix, as illustrated by the statements

>> num = det(c)
um = -35

Evaluate the determinant of \( c \).

>> tr = trace(c)
tr = 6

Evaluate the trace (sum of diagonal elements) of \( c \).

>> tran = c'
tran = 1 4
  10 5

Evaluate the transpose of \( c \). The special OCTAVE symbol \( ' \) is an operator for matrix transposition. The function transpose serves the same function.

>> inverse = inv(c)
inverse = -0.142857 0.285714
  0.114286 -0.028571

Evaluate the inverse of \( c \).

In OCTAVE, if we (matrix) multiply a row vector into a column vector, we obtain a number that is called the dot product of the \( n \)-dimensional vectors. That dot product can be quickly evaluated with statements such as

>> a = [1,2,3,4,5];
>> b = [6,7,8,9,10];
>> dotprod = a * b'
dotprod = 130

Create two vectors, suppressing display.

Evaluate dot product. Actually, OCTAVE has a function for evaluating dot products; the statement \( \text{dotprod} = \text{dot}(a,b) \) would have produced the same result.
Since the two vectors $a$ and $b$ are created as row vectors, the second factor in the dot product must be transposed to convert it into a column vector before the matrix product is evaluated. As one might expect, OCTAVE also has the analogous function \texttt{cross} for cross products, though that operation is meaningful only for vectors with three components.

### 4.3.4 Solving Linear Equations

OCTAVE’s matrix operations provide a quick route to solve simultaneous linear equations. Suppose, for example, we had the equations

\[
\begin{align*}
2x_1 + 5x_2 + 3x_3 &= 3 \\
-x_1 + 3x_2 - 4x_3 &= 1 \\
x_1 - x_2 &= 1
\end{align*}
\]

where the matrix expression uses the usual convention for matrix multiplication. We would enter and verify the coefficient matrix and the vector of inhomogenities with the statements

```matlab
>> a = [ 2.0,5.0,3.0; -1.0,3.0,-4.0; 1.0,-1.0,0.0 ]
a = 2 5 3
   -1 3 -4
   1 -1 0
>> b = [ 3.0; -4.0; 1.0 ]
b = 3
   -4
   1
```

Then, recognizing in the abstract that $ax = b$ implies that $x = a^{-1}b$, we might—though, from a numerical perspective, this approach is not recommended for any but the smallest systems—seek a solution with the statements

```matlab
>> ainv = inv(a);
>> x = ainv * b
x = 0.85294
   -0.14706
   0.67647
```

This solution correctly solves the original equations, as hand substitution—and the statement \texttt{[a*x, b]}—will confirm.

Actually, OCTAVE includes another operator that further facilitates the task of solving simultaneous linear equations. The solution $a^{-1}b$ could be seen as the end result of dividing the vector $b$ by the matrix $a$ from the left side. OCTAVE uses the operator \texttt{\} to convey this meaning. Thus, once $a$ and $b$ have been defined as above, the single statement

```matlab
>> x = a\b
x = 0.85294
   -0.14706
   0.67647
```

yields the same result.

To carry this exposition just one step further, we might have written the equations of interest in the alternative form

\[
\begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} 2 & -1 & 1 \\
5 & 3 & -1 \\
3 & -4 & 0 \end{pmatrix} = \begin{pmatrix} 3 \\
-4 \\
1 \end{pmatrix}
\]
4.3. A SAMPLING OF OCTAVE CAPABILITIES

i.e., in the form \( xC = b \), where \( x \) and \( b \) are now row vectors (and \( C \)—notice—is the transpose of \( A \)). In this case, the solution would be found by evaluating \( x = bC^{-1} \), i.e. by multiplying \( b \) from the right by the inverse of \( C \) (or dividing \( b \) from the right by \( C \)). The OCTAVE statements

\[
\begin{align*}
\text{>> } c & = [ 2,-1,1; 5,3,-1; 3,-4,0 ]; \\
\text{>> } b & = [ 3, -4, 1 ]; \\
\text{>> } x & = b/c \\
\end{align*}
\]

yield the result we have already obtained, though the vectors are now row vectors rather than column vectors. We could have also used \( c = a' \); (or \( c = \text{transpose}(a) \)) and \( b = b' \); or \( b = \text{transpose}(b) \)) to get the requisite transposed matrices.

4.3.5 A First Graph

OCTAVE’s commands also facilitate the graphing of known univariate functions. First, we need a vector of values—usually, but not necessarily, equally spaced—spanning the domain of the independent variable from some starting point to some ending point. In the first instance, OCTAVE uses the colon operator ‘:’ for this purpose. For example, we might create simple vectors with the statements

\[
\begin{align*}
\text{>> } y & = [ 0 : 1 : 5 ]; \\
\text{>> } yy & = [ -3 : 2 : 6 ]; \\
\end{align*}
\]

More to the point in preparing to graph a function of a variable \( x \), we might use a statement like

\[
\begin{align*}
\text{>> } x & = [ 0.0 : 0.1 : 10.0 ]; \\
\end{align*}
\]

or we might create the same vector using the OCTAVE function \texttt{linspace} in a statement like

\[
\begin{align*}
\text{>> } x & = \text{linspace}( 0.0, 10.0, 101 ); \\
\end{align*}
\]

Here, the arguments are, in order, the starting value, the ending value, and the number of points to be equally distributed in the interval between the starting and the ending values, \textit{including} the two bounding points. Thus, the number of segments into which the interval is divided is one \textit{less} than the value of the third argument.

If, for example, we want a graph of the hyperbolic cosine function using 100 divisions over the interval \(-3.0 \leq x \leq 3.0\), we might use the OCTAVE statements

\[
\begin{align*}
\text{>> } dx & = 6.0 / 100.0; \\
\end{align*}
\]

Determine segment size when a line of length 6 is divided into 100 segments.
Create a vector of 101 equally spaced values ranging from −3.0 to 3.0. We could also have used the statement
\[ x = \text{linspace}( -3.0, 3.0, 101); \]
Assign to each element of a vector \( y \) the hyperbolic cosine of the corresponding element of \( x \).

To draw the graph, OCTAVE passes the command to a subsidiary plotting toolkit. At least three such toolkits (\texttt{qt}, \texttt{gnuplot}, and \texttt{fltk}) are available in the standard distribution of OCTAVE. Before graphing anything, we need to find out what toolkits are available and select one, which is achieved by the statements

\begin{verbatim}
>> graphics_toolkit()
ans = qt
>> available_graphics_toolkits
ans =
{ [1,1] = fltk
  [1,2] = gnuplot
  [1,3] = qt }
>> graphics_toolkit( "gnuplot" )
\end{verbatim}

Select \texttt{gnuplot}, particularly if you ultimately want to incorporate Greek letters in PostScript files that are output.\footnote{See Section 4.17.1 for ways to incorporate Greek letters; see Section 4.12 for ways to output PostScript and PDF files. Note also (1) that either single or double quotation marks can be used and (2) that, once a toolkit has been selected, that choice remains in effect until the session is ended or until it is explicitly changed.}

\begin{verbatim}
>> graphics_toolkit( 'qt' )
\end{verbatim}

Restore default, which we will use unless we need Greek letters.

After selecting a graphics toolkit, we can invoke the \texttt{plot} command with the single statement

\begin{verbatim}
>> plot( x, y )
\end{verbatim}

Plot \( y \) versus \( x \), drawing straight line segments to connect the points whose coordinates are provided in the two vectors. Note that the vector containing values of the independent variable is the first argument.

The last command invokes the selected graphics toolkit, and opens a new window, the \textit{Figure} window, which shows the graph created by these statements. By default, the axes are drawn as a border around the graph and the graph itself is rendered in blue. A copy of the graph produced is shown in Fig. 4.1.

Taking control of some of the defaults adopted by the command \texttt{plot} or otherwise embellishing the graph can be accomplished in several ways. We might, for example,

- use the command \texttt{axis} with a four-component vector as argument to specify the desired ranges \((x_{\text{min}}, x_{\text{max}})\) on the \textit{x} axis and \((y_{\text{min}}, y_{\text{max}})\) on the \textit{y} axis.
- use the command \texttt{title} to place a title on the graph.
- use the commands \texttt{xlabel} and \texttt{ylabel} to put labels on the \textit{x} and \textit{y} axes.
- use the command \texttt{grid} to replace short tics along the axes with full grid lines.
- use the command \texttt{set}, which will set properties controlling the characteristics of a component of a graph, and the command \texttt{gca}, which specifically selects the axes in the current graph as the operative component.\footnote{Specifically, \texttt{gca} returns a handle—see Section 4.4—identifying the axes in the current graph.}
4.3. A SAMPLING OF OCTAVE CAPABILITIES

Figure 4.1: The hyperbolic cosine function.

- add keywords (properties—see Section 4.4) and values to the argument of the plot and set commands. For example, the keyword
  
  - "color", which admits an RGB triplet and also the color names "blue" (the default), "black", "cyan", "green", "magenta", "red", "white", and "yellow".
  - "linestyle", which admits the values "-" (solid line, the default), "--" (dashed line), ":" (dotted line), "-.
  - "linewidth", which is a number specifying the width of the line, with default 1, with 2 being twice as bold as the default, etc.
  - "fontsize", which is a number specifying the size of the font in points, with default 10 points.

Note, here, that OCTAVE automatically modifies plots. For example, after using one or more of the above commands, the existing plot will change automatically to reflect the execution of the command. Exploiting the additional commands and properties described in the previous paragraph, we might therefore produce the more readable and useful graph shown in Fig. 4.2 with the statements\textsuperscript{11}

\begin{verbatim}
>> plot( x, y, 'color', 'black', 'linewidth', 3.0 )
>> axis( [ -4.0, 4.0, 0.0, 12.0 ] )
>> grid on
>> title( 'Hyperbolic Cosine Function', 'fontsize', 20 )
>> xlabel( 'x', 'fontsize', 16 )
>> ylabel( 'cosh(x)', 'fontsize', 16 )
>> set( gca, 'fontsize', 16 )
\end{verbatim}

Here, we have invoked the property \texttt{fontsize} to specify the size of the fonts to be used for the title

\textsuperscript{11}Note that either single or double quotation marks can be used to enclose both property names and text values.
and labels. Note also (1) that the special characters `\n` in a text string will specify the start of a new line (a la coding in C) and (2) the statement `grid off` will suppress the display of the grid and the statement `grid` alone will toggle to the other state, e.g., from `off` to `on`.

Note, incidentally, the power of the statement `y = cosh(x)`. The argument of the function `cosh` is a vector and contains several values. In response to this statement, OCTAVE generates a second vector `y` having the same number of elements as `x`. Each element in `y` is the hyperbolic cosine of the corresponding element in `x`. In many languages, we would be obliged to use a more elaborate construction to instruct the computer to work on each element individually. OCTAVE automatically understands that intention with this simpler statement. Beyond simpler coding, OCTAVE also achieves faster execution than would be achieved in more traditional coding.

### 4.4 Properties, Objects, and Handles

In the example in Section 4.3.5, we encountered the notion of properties. From the perspective of object-oriented programming, the creation of a graph in OCTAVE entails the creation of several objects and the placement of these objects at appropriate points on the screen. The command `plot`, for example, first creates a `figure` object, displayed on the screen as the `Figure` window. Then, within the figure object—i.e., as a child of the figure object—`plot` creates an `axes` object, within which it creates a `line` object—the graph itself. The commands `title`, `xlabel`, and `ylabel` create `text` objects as children of the `axes` object. Each of these objects has numerous properties, all of which have default values and many of which can be changed with appropriate individual statements or equivalent inclusions in the full argument of one or another command. As here exploited, the

---

12 The property `fontunits` allows alternative units of measure to be stipulated.
13 Property names and text values for properties are insensitive to case, so we might for example alternatively use `FontSize` and/or `LineWidth` as a way to render the property names more legible.
14 Warning: PostScript files produced—see Section 4.12—from on screen displays may not incorporate these on-screen embellishments accurately.
15 Actually, the command uses the existing current figure object if there is one and creates a new one only if there is no existing one.
properties color, linewidth, and fontsize are among those which the user can change in this way.\textsuperscript{16}

With the substance of the previous paragraph in mind, we could create the graph in Fig. 4.2 much more deliberately with the statements

\begin{verbatim}
>> figure
>> axes('fontsize', 14, 'xlim', [-4.0,4.0], 'ylim', [0.0,12.0])
>> line( x, y, 'color', 'black', 'linewidth', 4 )
>> grid on
>> title( 'Hyperbolic Cosine Function', 'fontsize', 20 )
>> xlabel( 'x', 'fontsize', 16 )
>> ylabel( 'cosh(x)', 'fontsize', 16 )
\end{verbatim}

Here, we have used the command \texttt{figure} to create the figure object, the command \texttt{axes} to create an axes object within the figure object, and the command \texttt{line} to place the graph itself within the axes. Further, we have exploited several OCTAVE properties to take control of some of the characteristics of the resulting display.

Yet another way to achieve control over the properties of a display is to exploit the \texttt{handles} OCTAVE assigns to the objects in the display. Normally, these handles are not displayed. We can, however, capture those handles for the figure object, the axes object, line objects, and text objects as they are created. The statements\textsuperscript{17}

\begin{verbatim}
>> hf = figure
>> ha = axes
>> hl = line( x, y )
>> htitle = title('Hyperbolic Cosine Function' )
\end{verbatim}

for example, create the various objects with default properties \textit{store the handles for these objects in the indicated variables}, and (if not terminated with a semicolon) display the value of the automatically generated handle for the corresponding object. Once the handles are available, statements like

\begin{verbatim}
>> get( hf )
>> get( ha )
>> get( hl )
>> get( htitle )
\end{verbatim}

will display all possible properties and their current values, and statements like

\begin{verbatim}
>> set( hf )
>> set( ha )
>> set( hl )
>> set( htitle )
\end{verbatim}

will reveal not only the default value (in braces) but also all legal values for each property. More significantly, statements like

\textsuperscript{16}In a departure from OCTAVE’s normal case-sensitive feature, the property names in the contexts of these commands are not case sensitive. While we did not exploit that feature in Section 4.3.5, occasionally for easy reading we will set the first letter of each new word in the property name in upper case. We will, however, index all entirely in lower case.

\textsuperscript{17}Arguments specifying properties can, of course, be included in these statements. We omit them for the sake of the current example—and so that the \texttt{get} commands used shortly will reveal the default values of the several properties.
allow manipulation of individual properties after the object has been created.

Sometimes, we will forget to save handles that we will subsequently need. OCTAVE’s commands gcf, which returns the handle of the current figure object, and gca, which returns the handle of the current axes object, are provided to help us recover some of those handles without having to recreate the entire display. Thus, for example, the statements

```
>> ha = gca
>> htitle = get( gca, 'Title' )
```

will retrieve the corresponding handles for the axes object and the text object containing the title, respectively, while the statements

```
>> set( gcf, 'Color', 'red' )
>> set( gca, 'FontSize', 14, 'XLim', [-4.0,4.0], 'YLim', [0.0,12.0], ...
      'XGrid', 'on', 'YGrid', 'on' )
```

will set the background (border) color for the current figure object to red and adjust several properties in the current axes object without requiring us to know the handles of those objects. These functions are especially valuable when the display is created with plot, since plot returns only a handle of the line created after the figure and axes objects have been created.

### 4.5 Saving and Retrieving the OCTAVE Session

The command `save` in OCTAVE will save some or all of the variables in OCTAVE’s workspace. For example, the statement

```
>> save "example.oct" a b c
```

saves the variables `a`, `b`, and `c` in a file named `example.oct` in the default directory. If no variables are specified, then the command `save` saves all of the variables currently in the OCTAVE workspace.

Saved variables can be retrieved using the `load` command. To restore the file just saved, for example, we would use the statement

```
>> load "example.oct"
```

which loads into the OCTAVE workspace all of the variables saved in the file `example.oct` and gives each the name it had originally, overwriting any existing variables with the same names.

In the first instance, the function `save` saves OCTAVE variables in ASCII text files. Sometimes we may want to save the data in a more compact binary form, a task that can be accomplished using a switch. For example, the statements

```
18 The default directory is also referred to as the current directory.
19 See Section 4.18.4 for a fuller discussion of this directory and of means to change it.
20 If the default directory is unclear or ill-defined, specification of the full path to the file may be needed.
```
4.5. SAVING AND RETRIEVING THE OCTAVE SESSION

```octave
>> save -binary "example.bin" a b c
>> save -mat-binary "example.mat" a b c
```

will save the values of the variables \( a \), \( b \), and \( c \) in OCTAVE compact binary format and in the MATLAB Version 6 compact binary format, respectively. The variables to be saved are separated by spaces and come after the filename specification. To restore the files saved in this way, we would use the statements

```octave
>> load "example.bin"
>> load "example.mat"
```

which load into the OCTAVE workspace all of the variables saved in the file `example.bin` or `example.mat` and gives each the name it had originally, overwriting any variables with the same name.

At a lower level of coding, OCTAVE also makes available several commands for working with files directly. For example, suppose we had a \( 5 \times 5 \) array of integers stored in a variable \( a \). Then, the statements

```octave
>> id = fopen( 'textexample.dat', 'w' );
>> fprintf( id, '%5d %5d %5d %5d %5d
', a );
>> status = fclose( id );
```

will write the values in that variable as ASCII text. First, the command `fopen` opens a file in the default directory for writing (`'w'`), gives it the name `textexample.dat`, prepares it to receive data, and associates with it a unique Logical Unit Number (LUN), which—through the variable `id`—is used in subsequent references to this file. The variable `a` is then written to the file with the command `fprintf`, whose arguments specify the file, the format—see next paragraph—to be used in structuring each line of output, and the variable(s) to output. Finally, the file associated with the assigned `id` is closed.

Format specification is achieved with a string that resembles what would be used in C. In the string above, the specification `%5d` indicates that the value should be output as an integer (d-format) right-justified (default) in a field 5 characters wide. In addition, we have stipulated display of five values per line by inserting the new line character `\n` after the fifth item. The spaces in the format string stipulate that each five-character field in the line should be separated from the following field with two spaces. Further, the string specifying the format is scanned repeatedly until all values in \( a \) have been output.

The resulting ASCII text file may then be read by any application that respects the structure OCTAVE gives to the file. Reading the file into OCTAVE at some later date cannot, however, be done with the `load` command. Instead, the more elaborate process with the statements

```octave
>> b = zeros( 5, 5 );
>> id = fopen( 'textexample.dat', 'r' );
>> b = fscanf( id, '%5d', [5,5] );
>> status = fclose( id );
```

Create array of proper size.
Open file for reading.
Read data as a \( 5 \times 5 \) array into \( b \).
Close file.

---

21 Alternatives to `'w'` are `'r'` for reading an existing file and `'a'` for appending to an existing file.
22 Had the file contained floating point values, we would have needed other specifiers. The specifier `%n.mf` formats the number in ordinary decimal notation and right-justifies it in a field \( n \) characters wide with \( m \) digits after the decimal point, and the specifier `%n.me` formats the number in scientific notation (E-format) and right-justifies it in a field \( n \) characters wide with \( m \) digits after the decimal point. The formats `%d`, `%f`, and `%e` give OCTAVE authority to pick the finer details as appropriate to each number. Clearly, we need to know something about the nature of the values to be written to the file before a sensible format specifier can be constructed. Further information is available in Section 4.7 and in the OCTAVE manuals.
The resulting array $b$ will be the same as the original array that was written into the file. Note, incidentally, that the format string \%5d will be read repeatedly until all 25 items of data have been read; there is no need to structure a more elaborate format string.\textsuperscript{23}

One important warning about the structure of the file is necessary. For the example in this section, we assumed that we had a $5 \times 5$ matrix. In writing the elements of that matrix into the specified file, OCTAVE will step down the first column of the matrix, which contains five values, writing those values one at a time into the first row in the file. Then, OCTAVE will step down the second column, writing its five values in turn into the second row of the file. The process continues until the last column has been placed in the last row of the file. In effect, the matrix in the file is the transpose of the matrix in OCTAVE’s workspace. When this file is read back into OCTAVE by the procedure in the previous paragraph, this transposition will be undone and the array in OCTAVE will be identical to the original array, though (here) with a new name. If read into some other tool, care may be necessary on this front.

Please also be aware that detailed formatting of input and output is a very complicated task. Full information is in the OCTAVE manuals.

A final means to write OCTAVE output to a file uses the command \texttt{diary}, which saves to a file everything that is printed in the OCTAVE command window. The one argument to this command names the file and will be a string. For example, to save a session in which we illustrate the differences among the several matrix products, we might execute the statements\textsuperscript{24}

\begin{verbatim}
>> diary matrix_diary.dat
[or diary( 'matrix_diary.dat' )] Open diary file in default directory.
>> a = [ 1,3,0; 3,7,2; 0,2,4 ]
a =
  1   3   0
  3   7   2
  0   2   4
>> b = [ 1,2,3; 4,5,6; 7,8,9 ]
b =
  1   2   3
  4   5   6
  7   8   9
>> c = a .* b
  c =
  1   6
 12  35  12
 0  16  36
>> d = a * b
  d =
  13  17  21
 45  57  69
 36  42  48
>> diary off
[or diary( 'off' )] Close the file.
\end{verbatim}

Once generated, the file produced by \texttt{diary} can easily be edited in a text editor or incorporated in other documents.

4.6 Loops, Logical Expressions, and Conditionals

Among the most ubiquitous programming structures is the loop, which provides a means by which a statement or block of statements can be executed some number of times, typically with small changes controlled by a loop index. Several such structures are available in OCTAVE. The simplest

\textsuperscript{23}When the file was written, since we wanted the data to appear in five rows of five each and we needed to include the new line stipulation, we did need the more elaborate format string.

\textsuperscript{24}Quotation marks are not needed around the file name, though they can be used without generating an error.
is the for loop. Storing the squares of the integers from 1 to 10 in a vector, for example, is readily accomplished with the statements\textsuperscript{25,26}

\begin{verbatim}
>> for i = 1:10
   x(i) = i^2;
end
>> x
x =  1   4   9  16  25  36  49  64  81 100
\end{verbatim}

The for/end statement (which logically is a single statement) steps through all values—integers from 1 through 10, inclusive—of the loop index \(i\), calculating their squares and storing each in the proper element of \(x\),\textsuperscript{27} and the statement \(x\) simply displays the end result.\textsuperscript{28} We make two further observations: (1) If the vector \(x\) already exists in OCTAVE’s workspace with more than 10 elements, this loop will simply install new values in the first ten elements of the existing array and will leave all other elements untouched; a safer procedure would be to invoke the statement \texttt{clear x} before executing the loop; (2) OCTAVE will also correctly execute the single-line presentation

\begin{verbatim}
>> for i = 1:10 x(i) = i^2; end
\end{verbatim}

though the semicolon could be replaced with a comma had we wanted to display the vector \(x\) as each element is added.

With an embellishment of the index, the loop can even be stepped by something other than one, as in the statements

\begin{verbatim}
>> clear x
>> for i = 1:2:10 x(i) = i^2; end
>> x
x =  1   0   9   0  25   0  49   0  81
\end{verbatim}

In this version, only elements 1, 3, 5, 7, and 9 are computed; elements 2, 4, 6, and 8 are assigned the values 0, and element 10 is not created, since the loop is not executed for that value of the index.

If we wanted not only to fill the vector \(x\) but also to see the values and the indices individually as they are computed, we would have to incorporate both the calculation and the display in the loop, i.e., we would have to place two statements in the loop. Since OCTAVE takes all statements between the keyword for and the keyword end to be in the loop, we need merely place both statements between those keywords, placing each on a new line or—alternatively—separating them on a single line with a comma (or a semicolon—if the output is to be suppressed). Taking the second option, we would achieve the desired objective with the statements

\begin{verbatim}
>> clear x
>> for i = 1:2:10 x(i) = i^2; [i, x(i)], end
\end{verbatim}

\texttt{ans} =  1   1
\texttt{ans} =  3   9

\textsuperscript{25}Note that use of \(i\) for the loop index will override OCTAVE’s default understanding of \(i\) as \(\sqrt{-1}\). The default meaning will be restored by the statement \texttt{clear i}.

\textsuperscript{26}The coding here illustrated requires OCTAVE to adjust the size of the array \(x\) with each added value as the loop unfolds, a process that slows the execution. A more efficient coding would preallocate space for the entire array with a statement like \(x = \texttt{zeros}(1,10)\), which creates the \(1 \times 10\) vector of zeros and eliminates the overhead of repeated extension of the array by one element.

\textsuperscript{27}Without the semicolon after the statement in the for loop, we would have been presented at each step with a display of the values stored in \(x\) at that step.

\textsuperscript{28}Alternatively, of course, we might have invoked the statements \texttt{y=1:10} followed by \(x=y.^\wedge2\), but these statements would not have illustrated the explicit construction of a loop.
In addition to the for loop, in which we must know ahead of time how many times the loop is to be executed, OCTAVE provides other loops that are capable of deciding when to stop by monitoring the progress of the loop. A while loop to accomplish one of the above tasks might be structured in the form

```octave
>> clear x
>> i = 0; while i < 10 i=i+1; x(i) = i^2; end
>> x
x = 1 4 9 16 25 36 49 64 81 100
```

To construct this while loop, the index i is first initialized to zero. Then, as each pass through the loop begins, that index is tested to see whether it has yet reached the value 10, at which point the loop terminates. Because the index is incremented by one with each pass through the loop, ultimately the loop will, in fact, terminate.

In constructing the while loop in the previous paragraph, we used OCTAVE's operator < (less than) to express our first logical condition. OCTAVE, of course, possesses the standard six such operators: < (less than), > (greater than), == (equal to), ~= or != (not equal to), <= (less than or equal to), and >= (greater than or equal to). In addition, the operators ~ and ! toggle logical values between 'true' and 'false', so, for example, the condition i < 10 is equivalent to the condition ~(i >= 10). Finally, Boolean algebra on logical expressions is facilitated by the operators && (logical and) and || (logical or). OCTAVE also has the operators & and | for the complementary element-by-element logical conditions and the functions and, or, not, and xor.

Within OCTAVE, logical conditions have values and, if set to a variable, each such condition will create a 1 × 1 array of class logical. Only two values, however, are allowed: 1 for true and 0 for false. The statements

```octave
>> i = 5;
>> v = [ i < 10, i == 10, i > 10 ]
v = 1 0 0
```

(v = [ true, false, false ]) reveal those possibilities.

Logical conditions appear not only in controlling loops but also in structuring branches in a sequence of statements. As with most programming languages, OCTAVE also possesses if/else/end and if/else/elseif/end constructs, though the else and elseif clauses can be omitted if they are not required. Thus, for example, the statement

```octave
>> for i = 1:10
    if x(i) < 0.0
        x(i) = -x(i);
    end
end
```

will replace each negative element in a vector x with the corresponding positive value and the statement

```octave
>> if a > 0
    b = a
else
    b = -a
end
```
4.7 READING DATA FROM A FILE

will set \( b \) equal to the absolute value of \( a \) (though the function \texttt{abs} will do so more easily). As in the previous paragraphs, we could—with careful use of commas or semicolons—compact these statements onto fewer lines than here used.

4.7 Reading Data from a File

Frequently, we need to read data into OCTAVE from an appropriately structured ASCII file, which may have been written by another program or created with a text editor. Procedures for reading such files into OCTAVE are described in this section. Before the data can be read, OCTAVE needs information about the type of data and the size and number of arrays in the file. If the file has any headers, OCTAVE needs to be told where to start reading the \textit{real} data.

We illustrate the process with the ASCII file \texttt{radio.dat}, which was created by a FORTRAN program that simulates the decay of a radioactive material into a material that is itself unstable. If this file simply contained a single array of numbers, we could read it with the command \texttt{load} as described in Section 4.5. Unfortunately, the file begins with a single line containing labels as text and continues with 201 additional lines, each containing four numerical values, each separated from the next with one or more spaces.\textsuperscript{29} The first value in each line is a time, and the second, third, and fourth values are the quantities of the initial material, the intermediate material, and the final (stable) material, respectively, at that time.\textsuperscript{30} We read the data from this file with the statements

\begin{verbatim}
>> id = fopen('radio.dat', 'r');  \quad \text{Open the file for reading, assign id.}
>> ln = fgetl(id); \quad \text{Read the first line from file.}
>> quantity = fscanf( id, '%f', [4,201] ); \quad \text{Read data. The } 4 \times 201 \text{ array } quantity \text{ will be created and expanded as the data are read.}
>> status = fclose(id); \quad \text{Close file.}
>> quantity = quantity'; \quad \text{Transpose to } 201 \times 4 \text{ array.}
\end{verbatim}

The first line of the file is now stored in the (string) variable \texttt{ln} (and we shall pay no further attention to it), and, with the transposition,\textsuperscript{31} the numerical values in the file are stored in the 201 row \( \times \) 4 column array \texttt{quantity}. We might, for example, plot the time evolution of the quantities of the three materials with the statements

\begin{verbatim}
>> plot( quantity(:,1), quantity(:,2), 'color', 'black', 'linewidth', 3 )
>> hold on
>> plot( quantity(:,1), quantity(:,3), 'color', 'black', 'linewidth', 3 )
>> plot( quantity(:,1), quantity(:,4), 'color', 'black', 'linewidth', 3 )
>> text( 2.5, 850.0, 'A', 'fontsize', 16 )
>> text( 2.5, 300.0, 'B', 'fontsize', 16 )
>> text( 40.0, 800.0, 'C', 'fontsize', 16 )
>> set( gca, 'fontsize', 16 )
>> hold off
\end{verbatim}

The first of these statements selects the graphics toolkit. The second plots the quantity of the initial material (\texttt{quantity(:,2)}, the second column of the array) versus time (\texttt{quantity(:,1)}, the first column of the array). Since we executed the statement \texttt{hold on}, the fourth and fifth statements \textit{overplot} graphs of the quantities of the second and third materials versus time. Finally, the statements invoking the command \texttt{text}, in which locations are expressed in the coordinates of the graph, place the labels \texttt{A}, \texttt{B}, and \texttt{C} at appropriate points on the graph. Several new features of OCTAVE have been introduced in this example:

\textsuperscript{29}Commas or tabs can also be used to separate values in a line.
\textsuperscript{30}Do not fail to copy this file into your directory and examine its structure with a text editor.
\textsuperscript{31}Remember that OCTAVE will read the \textit{rows} of the file into the \textit{columns} of the array.
CHAPTER 4. INTRODUCTION TO OCTAVE

Figure 4.3: Simulation of radioactive decay.

- The command `hold on`, which suppresses the automatic erasing of the *Figure* window so that new invocations of the command `plot` will add to rather than replace the existing contents of that window. This feature is toggled the other way with the command `hold off`, while the command `hold` by itself toggles the feature to whatever is its other state.

- The command `text`, whose arguments use the units of the graph to specify the (x,y) location of the lower left corner of the text, the string to be positioned at the specified point, and may include a number of properties to control the style of the displayed text. Further, we have specified 16 point type, overriding the 10 point default.

- A strategy for “reading through” unwanted data in a file in the process of extracting the needed data from the file. In the above example, we recognize the presence of the initial labeling line and read past it.

The resulting plot is shown in Fig. 4.3.

Before leaving this example, we comment on additional features of the command `plot`. Among other useful features (see the OCTAVE manuals for full details), the command will accept more than one set of dependent variables. Thus, after the array `quantity` has been created as above, we might then execute the statements

```octave
>> t = quantity( :, 1 );
>> abc = quantity( :, 2:4 );
>> plot( t, abc )
```

In this form, OCTAVE will produce a similar figure to the one in Figure 4.3 (with no labels) without having to use the `hold` command and separate `plot` commands. In the absence of appropriate keywords in the `plot` command, this route will also use different colors for each of the three graphs.
4.8 On-line Help

OCTAVE provides a way for users to request help while in an OCTAVE session. As we have already seen, we can obtain information about the current status of the OCTAVE workspace or specific information regarding one or more of the variables stored in that workspace with the commands `who` and `whos`, perhaps with additional arguments specifying the particular variables about which information is desired. The statement

\[ \texttt{help function} \]

e.g., `help plot`, `help axis`, and `help meshgrid`, will display detailed information on the specified command directly in the GUI command window or in the CLI Shell window. In many cases the documentation more than fills the available screen. In some versions of OCTAVE, the entire help message is displayed. In other versions, only a screen full will be displayed at a time, and typing `f` or the space bar will advance to the next screen, typing `b` will go back to the previous screen, and typing `q` will exit the help message and return to the OCTAVE prompt.

In addition, the entirety of OCTAVE’s documentation is available online at their website [www.octave.org](http://www.octave.org), which can be accessed from an item in the Help menu in the GUI.

4.9 m-Files

Any command file called by OCTAVE is saved as an m-file, so named because the file itself by convention is named with a file type `.m`. A command file may be a function file, which defines a function or procedure controlled by arguments and effecting isolation of all variables internal to the function, or simply a batch execution file or script, which contains a succession of statements that we might alternatively type directly at the command line. OCTAVE makes available many built-in functions. We have already introduced several (`plot`, `fopen`, `inv`, `zeros`, `fix`, `sin`, `cosh`, `sqrt`, ...). Others include functions for drawing histograms, creating common statistical representations, producing colormaps and creating common matrices.

OCTAVE’s built-in functions can be supplemented with user-written scripts and functions. The script is the simpler of these options. We merely create a file containing exactly the statements we might otherwise type, one at a time, interactively. For example, we might create a file containing the lines listed in Table 4.1. Then, after this file has been stored with the name `plotradio.m` (say) in a directory that is in OCTAVE’s search path (see Section 4.18 and the Local Guide), the single statement

\[ \texttt{>> plotradio} \]

will instruct OCTAVE to execute each of these statements in turn and ultimately produce the graph we have already shown in Fig. 4.3. Especially when we are doing several similar tasks or developing and debugging an extended sequence of statements, the creation of a script, which can be easily edited, saved, and then reexecuted, can be an immense time saver. Indeed, the ‘Editor’ tab in the GUI facilitates this process.

We can also define customized procedures and functions. The reserved word `function` on the first non-blank line of a file informs OCTAVE that a user-defined function follows. As with scripts, every user-defined function file should be named with the file type `.m`. For example, the statements listed in Table 4.2 define a function that plots the function `funct` (which must

---

32 When no file type is specified, the file type `.m` is assumed by default.

33 Since the function `luplot` assigns no value to the variable `fct`, we could have omitted that variable (and the following equal sign) altogether in the first line of the file defining the function or, alternatively we could have conveyed the absence of a returned value by replacing `fct =` with `[ ]`.  
34 See the last paragraph of this section for a discussion of the significance of the comments before the first executable statement in this file.
CHAPTER 4. INTRODUCTION TO OCTAVE

Table 4.1: The OCTAVE script plotradio.m. Comments introduced by a number sign can be included in the file; everything from the number sign to the end of the line will be ignored when the line is executed by OCTAVE.)

```octave
# plotradio.m

id = fopen('radio.dat', 'r');      # Open the file for reading,
    # assigning id
ln = fgetl(id);                   # Read first line from file
quantity = fscanf(id, '%f', [4,201]); # Read data into array
status = fclose(id);              # Close file
quantity = quantity';             # Transpose array
graphics_toolkit( 'gnuplot' );    # Select graphics toolkit
plot(quantity(:,1), quantity(:,2)) # Plot A vs t
    hold on
plot(quantity(:,1), quantity(:,3)) # Add B vs t
plot(quantity(:,1), quantity(:,4)) # Add C vs t
text( 2.5, 850.0, 'A', 'fontsize', 16 ) # Label A
text( 2.5, 300.0, 'B', 'fontsize', 16 ) # Label B
text( 40.0, 800.0, 'C', 'fontsize', 16 ) # Label C
hold off
```

Table 4.2: The OCTAVE function luplot.m.

```octave
function fct = luplot(funct, N, start, stop)
    # LUPLOT - Plots user-specified function.
    # Function luplot is passed a function identified by the
    # (string) variable funct and then, in order, the number of
    # segments into which the interval is to be divided, the
    # starting value of x, and the stopping value of x for the plot.
    # It returns a plot of the function but assigns no value to the
    # variable fct.

dx = ( stop - start ) / N;     # Calculate increment
x = [ start : dx : stop ];     # Create vector of independent variable
y = feval(funct,x);            # Evaluate dependent variable
plot(x, y, "color", "black", "linewidth", 2 );
    # Plot graph
set( gca, 'fontsize', 16 );   # Set size of labels
endfunction
```

be enclosed in quotation marks), evaluating it at \(N+1\) equally spaced points spanning the interval \(\text{start} \leq x \leq \text{stop}\). The function itself is invoked by typing luplot followed by the requisite arguments in parentheses. For example, if luplot is stored in a directory in OCTAVE’s search path,\(^{35}\) either of the statements\(^{36}\)

\[^{35}\text{See Section 4.18.}\]
\[^{36}\text{The symbol @ in this context creates a handle for the function whose name follows the symbol, which handle then points to the function. The older notation ’cos’ still works but is now regarded as obsolete. In these and similar contexts, we shall use the @ sign throughout.}\]
Table 4.3: The OCTAVE function lineshape.m.

```octave
function y = lineshape(x)
# LINESHAPE - Plots Lorentz line shape.
# The function lineshape returns the value of the Lorentz
# distribution for given argument x. The parameters a, b, and
# x0 are explicitly coded in the function.

a = 1.0;
b = 1.0;
x0 = 2.0;
y = a^2./(b^2 + (x-x0).^2);
endfunction
```

```
>> luplot( @cos, 200, 0.0, 10.0 )
>> luplot( 'cos', 200, 0.0, 10.0 )
```

will generate a graph of the cosine function \( \cos(x) \) over the interval \( 0 \leq x \leq 10 \), dividing the interval into 200 segments (i.e., plotting 201 points uniformly distributed over the interval). Indeed, the pair of statements

```
>> g = @cos
>> luplot( g, 200, 0.0, 10.0 )
```

which assign the function handle to the variable \( g \) which is then used in the argument to \( \text{luplot} \) will also generate the graph.

What if, however, we wanted to use \( \text{luplot} \) to graph a function that is not built into OCTAVE? In that case we would first write a \textit{function} m-file defining the function of interest. For example, suppose we wanted a graph of the Lorentz lineshape defined by

\[
y(x) = \frac{a^2}{b^2 + (x-x_0)^2}
\]

for particular values of \( a \), \( b \), and \( x_0 \). We would begin by writing a \textit{function} m-file defining the Lorentz lineshape, perhaps using the coding listed in Table 4.3. Then, after the file \textit{lineshape.m} has been stored somewhere in OCTAVE's search path,

```
>> luplot( @lineshape, 200, -10.0, 10.0 )
```

to call \( \text{luplot} \). Momentarily, the graph of the Lorentz lineshape shown in Fig. 4.4 will appear on the screen.

With the approach of the previous paragraph, we would have to edit the function m-file if we wanted to graph the Lorentz line shape for a different set of parameters. Because of the way \( \text{luplot} \) is written, however, we cannot simply add these parameters as arguments to \textit{lineshape}. Were we to adopt that approach, we would want to set the parameters \( a \), \( b \), and \( x_0 \) at OCTAVE's command level and have them known not within \( \text{luplot} \) itself but within a function that \text{luplot} calls. The statement \textit{feval} in \textit{luplot} does not include the passing of parameters to \textit{lineshape}. To achieve this passing of a parameter from OCTAVE's command level into a function that is not directly called

\[37\]See Section 4.18.4.
from the command level, we elect to make use of global variables.\footnote{The risks associated with global variables can be avoided by using the function \texttt{varargin}, whose use, however, is involved. See the OCTAVE manuals for full detail.} In essence, at the command level, we declare appropriate variables—here a, b, and x0—to be global and give them the desired values. Then, we declare these same variables to be global in the function that is to access those values. In this way, we circumvent OCTAVE’s automatic isolation of variables in functions from variables of the same name in other functions. To this end, we might rewrite \texttt{lineshape} to be as listed in Table 4.4 and store it in the file \texttt{lineshape1.m}. Here, the statement \texttt{global A B X0} tells the function that it is to find the values of the variables A, B, and X0 in variables named A, B, and X0, which will have been declared \texttt{global} and assigned values at command level.\footnote{We will consistently use upper-case letters for the global variables. Since we routinely use lower-case letters for ordinary variables, this convention reduces the risk of inadvertent use of global variables in inappropriate contexts.} Then, we invoke \texttt{luplot} with \textit{this} function using the statements

\begin{verbatim}
function y = lineshape1(x)
  # LINESHAPE1 - Plots Lorentz line shape.
  # The function lineshape returns the value of the Lorentz
  # distribution for given argument x. The parameters A, B, and
  # X0 must be declared global at the command line and assigned values
  # prior to invoking lineshape1, which then uses the values in those
  # global variables.
  global A B X0
  y = A^2./(B^2 + (x-X0).^2);
endfunction
\end{verbatim}
4.10. EIGENVALUES AND EIGENVECTORS

 Declare global variables.
 Set parameters.
 Graph function.

 and produce again the graph shown in Fig. 4.4. This time, however, our approach is more flexible,
since we can more easily change the values of the parameters and produce another graph.

 The above m-files have made use of OCTAVE’s capacity for self-documentation. We have again
introduced comments with number signs. In function m-files, comment lines that precede the first
executable line in the file have special significance. The statement

 lookfor CharacterString

 searches the first comment line of every m-file that it finds and displays every such line containing
the specified character string. Further, the statement

 help FileName

 displays all comments preceding the first executable statement. The commands lookfor and help
will thus automatically include any user-added m-files in their scope, provided each such file is
structured with appropriate comments at its beginning. The wisdom of documenting user-written
functions thoroughly cannot be overstressed, and this particular feature of MATLAB makes it easy
to keep the documentation coordinated with—and in the same file as—the coding itself.

4.10 Eigenvalues and Eigenvectors

Eigenvalues and eigenvectors of symmetric matrices figure prominently in many physical contexts
(finding quantum mechanical energies and energy shifts, studying small amplitude oscillations, ...).
The OCTAVE function eig takes as arguments the name of the matrix whose eigenvalues and
eigenvectors are desired. The function returns either the eigenvalues or the eigenvectors and the
eigenvalues, depending on the structure of the variable to which its output is directed. To find
eigenvalues alone, we need only create the matrix and invoke eig with statements like

 Define matrix whose eigenvalues are sought
 and display it.

 Invoke eig to find eigenvalues.

 If, however, we want both the eigenvectors and the eigenvalues, we would use a statement like

 If, however, we want both the eigenvectors and the eigenvalues, we would use a statement like

\[
\begin{bmatrix}
1 & 3 & 0 \\
3 & 7 & 2 \\
0 & 2 & 4
\end{bmatrix}
\]

40 Alternatively, percent signs could be used.
41 Quotation marks around CharacterString are not necessary but can be included.
42 Quotation marks around FileName are not necessary but can be included.
With this statement, each column of the array \texttt{evecs} is one of the normalized eigenvectors of the matrix \texttt{A}; the eigenvalues are returned in a diagonal matrix.

The format of the output makes it easy to verify that the results are correct. We want the eigenvalues and eigenvectors to satisfy the simple equation

\[ A \mathbf{x} = \lambda \mathbf{x} \quad (4.5) \]

where \( \lambda \) is an eigenvalue and \( \mathbf{x} \) is the corresponding eigenvector. For the first eigenvalue and eigenvector, we find that

\[
\begin{align*}
\text{Extract first eigenvalue.} & \\
\text{Extract first eigenvector.} & \\
\text{Construct two-column array with left side of Eq. (4.5) as first column and right side as second column.}
\end{align*}
\]

\[
\text{ans} = \begin{bmatrix}
-0.346968 & -0.346968 \\
0.160683 & 0.160683 \\
-0.073215 & -0.073215
\end{bmatrix}
\]

Inspection shows that these results are the same. Similar statements focusing on the second and third eigenvalues will confirm those results as well.

Alternatively, we can verify all eigenvalues and eigenvectors at the same time. To obtain the left hand side of Eq. (4.5), we simply evaluate

\[
\begin{align*}
\text{To obtain the right hand side, however, we must recognize that the diagonal character of the matrix of eigenvalues means that we must evaluate the product in a seemingly unnatural order, specifically}
\end{align*}
\]

\[
\begin{align*}
\text{\texttt{A * evecs}} & \\
\text{\texttt{evecs * evals}}
\end{align*}
\]

\[
\begin{align*}
\text{\texttt{rhs}} & \\
\text{\texttt{lhs}}
\end{align*}
\]

Comparison of these two matrices confirms the correctness of all three eigenvalues and eigenvectors at once. Indeed, knowing that internal arithmetic is done in double precision, we might even examine the differences between these two sides with the statement

\[
\begin{align*}
\text{\texttt{rhs - lhs}}
\end{align*}
\]

where the premultiplier \( 10^{-15} \) applies to all values.\footnote{The precise form of this output may vary from version to version and platform to platform. In all cases, however, the values will be zero within computer roundoff.} Evidently, the two matrices, whose values are on the order of 1, differ by values on the order of \( 10^{-15} \)—surely equal within the roundoff errors endemic to (double precision) computer calculations.\footnote{Note that these last results are largely determined by roundoff errors within the computer and may well differ from computer to computer. In all cases, the results will be negligibly small.}

\textbf{OCTAVE's command} \texttt{eig} works on matrices that must, of course, be square, but need not be symmetric or real.
4.11 Graphing Scalar Functions of One Variable

The most common—and simplest—graph is a two-dimensional plot of a function of one variable, a graph of dependent variable versus independent variable. As we have already seen in Section 4.3.5, the basic command for producing such a graph is `plot`, which—in its simplest form—has the syntax

```matlab
>> plot( independent variable, dependent variable )
```

where the two arguments identify vectors in which values of the variables to be plotted are stored. The graph itself is constructed by drawing straight line segments connecting consecutive points identified in the vectors. Both vectors must, of course, have the same number of elements.

The action of `plot` can, however, be modified by stipulating values for one or more properties. We have already introduced a number of functions used to embellish graphs including `title`, `xlabel`, `ylabel`, `grid`, `hold`, and `text`. Other properties and functions will be introduced as the need for them arises.

4.11.1 The Basic Strategy

As a first example, suppose that we desire a graph of the magnetic field on the z axis of a circular current loop in the xy plane. As a function of position, this field is given in dimensionless form by

$$B(z) = \frac{1}{(1 + z^2)^{3/2}} \quad (4.6)$$

Suppose that we want a graph over the interval $-4.0 \leq z \leq 4.0$ with 100 divisions of the interval (101 points plotted). We would then execute the OCTAVE statements

```matlab
>> dz = 8.0/100.0;          % Set increment between points.
>> z = [ -4.0 : dz : 4.0 ]; % Create vector of 101 values in interval.
>> B = (1.0 + z.^2).^-1.5;  % Evaluate B, also a vector of 101 values.
>> plot( z, B, 'color', 'black', 'linewidth', 2 ) % Plot graph.
>> set( gca, 'fontsize', 14 ) % Enlarge tic labels.
>> title( 'Magnetic Field on z axis', 'fontsize', 20 ) % Title graph.
>> xlabel( 'Dimensionless Position', 'fontsize', 16 ) % Label horizontal axis.
>> ylabel( 'Magnetic Field', 'fontsize', 16 ) % Label vertical axis.
```

where the functions `title`, `xlabel`, and `ylabel` and the `fontsize` property have the effects described in Sections 4.3.5 and 4.12. This sequence of statements will produce the graph shown in Fig. 4.5.

4.11.2 Marking Points:

Sometimes, we may wish to mark the individual points in a graph with a distinct symbol. The alternative version

```matlab
>> plot( z, B, "color", "black", "." )
```

of the above `plot` command will plot a dot at each point and suppress the lines connecting these dots. A version explicitly using the `marker` property, i.e.,

```matlab
>> plot( z, B, "color", "black", "marker", "." )
```

will plot a dot at each point and also connect the dots with lines. The wide variety of available markers is shown in Table 4.5. Additional marker properties include...
4.11.3 Plotting Several Graphs on One Set of Axes

Many times we may want to place several different graphs on the same set of axes. The simplest way to achieve this end is to specify an array as the value of the second argument (dependent variable) to the command `plot`. We might, for example plot superimposed graphs of the sine and cosine functions with the statements
Figure 4.6: Undamped and damped sine waves.

Construct vector of values of independent variable.

Evaluate dependent variables.

OCTAVE is also capable of producing polar plots. For example, to graph the cardioid defined in polar coordinates by the equation

\[ r(\theta) = a(1 - \cos \theta) \]  

(4.7)
Figure 4.7: Polar plot of the cardioid. Note that versions of OCTAVE prior to Version 4.0 displayed the plot on Cartesian rather than cylindrical axes.

and shown in Fig. 4.7, we might execute the statements

\begin{verbatim}
>> dt = 2.0*pi/100.0;
>> theta = [ 0 : dt : 2*pi ];
>> a = 3.0;
>> r = a*(1.0 - cos(theta));
>> h1 = polar( theta, r );
>> set( h1, 'color', 'black', 'linewidth', 4 )
>> set( gca, 'fontsize', 16 )
\end{verbatim}

Create vector of values of independent variable (101 values in \(0 \leq \theta \leq 2\pi\)).
Set parameter.
Evaluate dependent variable.
Plot polar graph, saving pointer to graph
Set color and linewidth for graph

Here, in the command \texttt{polar}, the azimuthal coordinate—in radians—comes first and the radial coordinate second. Equal increments are automatically used on each axis so as to produce a figure that does not appear stretched in either direction. Further, because the command \texttt{polar} does not admit specification of many properties, we have jumped the gun and used the \texttt{set} command to modify properties of the display, though we then have to define \texttt{h1} as a "handle" (or pointer) so we can identify the graph in the \texttt{set} command.\textsuperscript{45}

\subsection*{4.11.5 Multiple Plots On a Page}

Sometimes we wish to plot several separate graphs in a single \textit{Figure} window. OCTAVE gives us substantial flexibility in formatting the layout of these graphs by providing the command \texttt{subplot}, which controls how many plots appear and where they appear in the \textit{Figure} window. The command has the general form \texttt{subplot(m,n,p)}, in which \(m\), \(n\), and \(p\) are integers with

- \(m\) specifying the number of plot \textit{rows} in the \textit{Figure} window,
- \(n\) specifying the number of plot \textit{columns} in the \textit{Figure} window, and

\textsuperscript{45}Handles and the \texttt{set} command are more fully discussed in Section 4.4.
• \( p \) specifying which subarea is the current active area. (The numbering of subareas starts with 1 in the upper left and migrates across the first row from left to right, then across the second row, etc.)

As a quick example using `subplot`, suppose we wish to plot the sine and cosine functions along with their analogous hyperbolic functions. We want a \( 2 \times 2 \) array of graphs, so we will use the statement `subplot(2,2,?)`. The plot shown in Fig. 4.8 can then be produced with the statements

```matlab
>> dz = 8.0/100.0;
>> z = [ -4.0 : dz : 4.0 ];
>> v = sin(z); w = cos(z); x=sinh(z); y=cosh(z);
>> subplot(2,2,1)
>> plot( z, v, 'color', 'black', 'linewidth', 2 )
>> set( gca, 'fontsize', 14 )
>> text(1.0, -0.5, 'Sine', 'fontsize', 14 )
>> subplot(2,2,2)
>> plot( z, w, 'color', 'black', 'linewidth', 2 )
>> set( gca, 'fontsize', 14 )
>> text( -1.0,-0.5, 'Cosine', 'fontsize', 14 )
>> subplot(2,2,3)
>> plot( z, x, 'color', 'black', 'linewidth', 2 )
>> set( gca, 'fontsize', 14 )
>> text( -3.5, 20, 'Hyperbolic Sine', 'fontsize', 14 )
>> subplot(2,2,4)
>> plot( z, y, 'color', 'black', 'linewidth', 2 )
>> set( gca, 'fontsize', 14 )
>> text( -3.0, 20.0, 'Hyperbolic Cosine', 'fontsize', 14 )
>> subplot(1,1,1)
```

Note that, after you have examined the graph, `subplot` should be reset with `subplot(1,1,1)` to specify one plot per plot window so as not to disrupt subsequent graphing.\(^{46}\)

### 4.11.6 Plotting Experimental Data

As a final example of two-dimensional graphing, we describe OCTAVE’s ability to produce plots of experimental data complete with error bars, representing each of the data points with a plotting symbol not connected with lines and using logarithmic scales on either (or both) axes. OCTAVE will also allow us to set the range of the axes. This example uses data from an experiment on an RC high-pass filter. Suppose the ASCII file `rcdata.dat` containing the data was created with a text editor and stored in the default directory or in OCTAVE’s search path. When printed, the file produces a two-column display, the first column containing the frequencies at which gains were measured and the second containing the measured gains. In all, the file contains twenty-one measurements (twenty-one lines). The data are read into the OCTAVE workspace with the statements

```matlab
>> id = fopen( 'rcdata.dat', 'r' );
Read data.
>> data = fscanf( id, '%f', [2,21] );
Close file.
>> fclose( id );
Transpose array.
```

Next, we plot the data using the statements

```matlab
>> semilogx( data(:,1), data(:,2), '.', 'color', 'black' )
>> axis( [ 10.0, 10000.0, 0.0, 1.1 ] );
```

\(^{46}\)The command `subplot` can also be written with a single three-character argument, e.g., `subplot(221)`, which is understood to be `subplot(2,2,1)`. Note that all three digits must be less than 10 for this format.
Figure 4.8: Graphs of sine, cosine, hyperbolic sine, and hyperbolic cosine on the interval $-4 \leq x \leq 4$.

>> title( 'Experimental Data for High Pass Filter', 'fontsize', 20 )
>> xlabel( 'Frequency', 'fontsize', 16 )
>> ylabel( 'Gain (Vout/Vin)', 'fontsize', 16 )
>> set( gca, 'fontsize', 16 )

These statements plot the values in the second column of the array data (gain measurements, the dependent variable) versus the values in the first column (frequency measurements, the independent variable), specifying that a logarithmic scale be used on the x axis and a linear scale on the y axis. Further, the stand-alone argument ‘.’ in the semilogx command results in a graph in which the points are marked with a bold dot but no line is drawn. The full graph, whose beginnings have been created by the above statements (and whose completion will be described in what follows), is presented in Fig. 4.9.

A graph of experimental data, however, is not complete without error bars. OCTAVE’s commands for making errorbars involve the statements

>> unc = data(:,2)*0.05;  % Compute +/- uncertainties at 5%.
>> hold on
>> errorbar( data(:,1), data(:,2), unc, '.k' );

After calculating the uncertainties, we invoke the OCTAVE command errorbar, providing the x and y coordinates, the uncertainty at each point, and the marker style to be used.

A graph of experimental data should also include a plot of predicted theoretical results. The

---

47 Three plotting combinations make use of logarithmic axes, the commands being semilogx for a logarithmic scale on the x axis only, semilogy for a logarithmic scale on the y axis only, and loglog for logarithmic scales on both axes.

48 See Section 4.11.2.

49 You can learn about the wide variety of arguments accepted by the errorbar command by examining what is displayed by the command help errorbar.
4.12. MAKING HARD COPY

The equation for the theoretical gain of this circuit is

\[
G = \frac{V_{\text{out}}}{V_{\text{in}}} = \frac{1}{\sqrt{1 + (f_0/f)^2}} \quad \text{where } f_0 \text{ is defined as } f_0 = \frac{1}{2\pi RC}
\]

(4.8)

Because we are using a logarithmic scale on the horizontal axis, however, we must calculate the vector of independent variables carefully. The variable \( f \) needs to have a range from \( 10^1 \) Hz to \( 10^4 \) Hz. For the most satisfactory plotting, we need to create a vector of values that will represent this range uniformly on a logarithmic scale. To achieve that objective, we begin by creating a vector—call it \( \text{temp} \)—containing a suitable number of values uniformly distributed on a linear scale from 1.0 to 4.0. Then, we take the values of \( f \) to be the base-10 exponential of the values in \( \text{temp} \), i.e.

\[
f = 10^{\text{temp}} = (e^{2.3026})^{\text{temp}} = e^{2.3026\text{temp}}
\]

(4.9)

The theoretical gain function is then evaluated using these values of \( f \). When the logarithm of the independent variable is taken to create the logarithmic scale, the plotted points will in fact be equally spaced on the logarithmic scale.

Having thus determined suitable values of \( f \), we evaluate the theoretical gain and overplot this function on the existing graph with the statements

\[
\begin{align*}
&>> r = 1.5e4; \quad \text{Set value of } R. \\
&>> c = 0.0442e-6; \quad \text{Set value of } C. \\
&>> f0 = 1/(2.0*pi*r*c); \quad \text{Calculate implied reference frequency.} \\
&>> \text{temp} = [1.0:0.12:4.0]; \quad \text{Create vector with values from 1.0 to 4.0.} \\
&>> f = \exp(2.3026 \times \text{temp}); \quad \text{Create frequencies from } 10^1 \text{ to } 10^4. \\
&>> \text{gtheory} = 1.0/\sqrt{1.0 + (f0^2/f.^2)}; \quad \text{Calculate theoretical curve.} \\
&>> \text{semilogx}(f, \text{gtheory}, \text{’color’, ’black’,} \\
&\quad \text{’linewidth’, 2}) \quad \text{Overplot theoretical curve.}
\end{align*}
\]

Here, we have used the values \( R = 15 \) k\( \Omega \) and \( C = 0.0442 \) \( \mu \)F and then overplotted the theoretical curve.

The statements listed above and used to produce Fig. 4.9 have been compiled into the file \texttt{plotrc.m}, which is available where all other files mentioned in this chapter are found.

4.12 Making Hard Copy

4.12.1 ... of Text

To print mathematical results from OCTAVE when OCTAVE is being run on an X-window device or in Windows, we can copy the desired text by merely highlighting it. Then, within a text editor, one can use the traditional pasting method (typically (Control-V)) to put the copied text into the text editor for later use, though we may wish to delete a number of unnecessary blank lines that may be introduced into the copied text. Alternatively, the command \texttt{diary} described in Section 4.5 will create an ASCII file containing everything written to the screen in an OCTAVE session, which can then be edited to extract only the portions wanted.

4.12.2 ... of Graphs

After a function or an array has been plotted on the screen, we sometimes wish to have a hard copy of the graph. Making that hard copy requires two steps. First, we must create a file containing an encapsulated PostScript description of the plot.\footnote{See also the function \texttt{logspace} described in the OCTAVE manuals.} To create that description, we

\footnote{Since many word processing and publishing packages have the capability to import PostScript files, this file is also useful when the graph is to be incorporated in a larger document.}
Figure 4.9: Data from the RC high-pass filter experiment.

- Create the graphical display in the *Figure* window.
- Make sure that the current directory points to the directory in which we wish the file to be deposited.\[52\]
- Execute either
  - the command `print -deps FileName.\text{eps}` which will create an encapsulated level 1 PostScript file containing the graphical display in the currently active window, giving that file the specified name and storing it in the current directory, or
  - the command `print -deps2 FileName.\text{eps}` which will create an encapsulated level 2 PostScript file containing the graphical display in the currently active window, giving that file the specified name and storing it in the current directory.\[53,54\]

Note that the current directory can be overridden by including an explicit path with the specified file name in the `print` command. Further, note that the `print` command admits numerous devices beyond `eps` and `-deps2` in the `-d` option. In particular, the devices `-deps` and `-deps2` will render colored displays automatically in a gray scale in the PostScript file; the devices `-depsc` and `-depsc2` will include color in the PostScript file; and the device `-dpdf` will output the graph to a PDF file. Finally, note that the command `help print` will display full information about the `print` command, including a comprehensive listing of all available devices.

Unfortunately, some features of a graph as it appears on the screen may not be exactly replicated in the output produced by a PostScript or PDF file. In particular the effect of the `color`, `linewidth`, and `fontsize` properties in the `plot`, `title`, `xlabel`, and `ylabel` commands may change when a graph is directed to a new output device. Thus, when these (and perhaps other) properties are specified, the values used to produce the desired plot on the screen may require adjustment to yield the desired plot when the file is printed.

\[52\] See Section 4.18.3 for ways to edit this current directory.
\[53\] Note that the `fltk` graphics toolkit always produces PostScript level 3.
\[54\] Single or double quotation marks around the file name are not necessary but can alternatively be used.
4.13 Graphing Scalar Functions of Two Variables

Suppose the function we wish to explore graphically is a function of two independent variables, for example,

\[ z(x, y) = \sin(2\pi x) \sin(3\pi y) \]  

(4.10)

which defines the shape of the [2, 3] mode of oscillation for a square membrane, or

\[ I(\xi, \eta) = \left( \frac{\sin \xi}{\xi} \right)^2 \left( \frac{\sin \eta}{\eta} \right)^2 \]  

(4.11)

which is related to the intensity in the diffraction pattern produced by a square aperture. In graphing a function of a single variable, we used as input to plot the values of the function at a selected set of (regularly or irregularly spaced) values spanning the desired range of that variable. Here, we need as input to any graphing routine the values of the function at a regular,\textsuperscript{55} two-dimensional grid of selected points \((x_{ij}, y_{ij})\) or \((\xi_{ij}, \eta_{ij})\) covering the portion of the \(xy\) or \(\xi\eta\) plane within which we seek to display the function graphically. In the broadest of terms, our task then will involve three steps:

1. Create two arrays, one \((x, x_i, \ldots)\) containing the first coordinate \((x, \xi, \ldots)\) of all the points in the grid and the other \((y, \text{eta}, \ldots)\) containing the second coordinate \((y, \eta, \ldots)\) of those points.

2. Evaluate the function over that grid, i.e., generate the array containing values of the dependent variable at the points identified in the arrays containing the independent variables. Because OCTAVE processes arrays element by element, a single statement, which would have the forms

\[
\begin{align*}
z &= \sin(2.0*\pi*x) .* \sin(3.0*\pi*y); \\
I &= (\sin(xi)./xi)^2 .* (\sin(eta)./eta)^2;
\end{align*}
\]

for the functions in Eqs. (4.10) and (4.11), will usually suffice.

3. Invoke one or another graphical display routine, specifying the array \((z, I, \ldots)\) containing the dependent variable as input and perhaps specifying one or more further arguments to control details of the display.

In this section, we describe how to create the necessary arrays containing values of the independent variables and then illustrate how to create different displays of these functions using mesh, surf, contour, and pol2cart.

4.13.1 A Preliminary: The Function meshgrid

In OCTAVE, the two-dimensional analog of the statement \(x = [\text{start} : \text{dx} : \text{stop}]\) is provided by the command meshgrid, which creates a pair of two-dimensional arrays, one containing the \(x\) coordinates of a grid of points uniformly spaced in a region of the \(xy\) plane and the other containing the \(y\) coordinates of points in that grid. Thus, for example, the statement

\[
\text{>> } [x,y] = \text{meshgrid}( -3.0:2.0:3.0, -1.0:0.5:1.0 );
\]

creates the arrays\textsuperscript{56}

\textsuperscript{55}With more than one independent variable, the use of irregular grids poses a particularly imposing challenge for creating the ultimate display. We elect to limit our discussion to a regular grid of values for the independent variable.

\textsuperscript{56}Note that the division of the \(x\) interval \([-3.0,3.0]\) into three segments results in four values in the array (i.e., \(-3, -1, 1, 3\)). In the same manner, the division of the \(y\) interval \([-1.0,1.0]\) into four segments results in five values \((-1, -0.5, 0, 0.5, 1)\).
>> x
x =
-3  -1   1   3
  -3  -1   1   3
   -3  -1   1   3
    -3  -1   1   3

>> y
y =
-1.0000 -1.0000 -1.0000 -1.0000
  -0.5000 -0.5000 -0.5000 -0.5000
   0.0000  0.0000  0.0000  0.0000
  0.5000  0.5000  0.5000  0.5000
  1.0000  1.0000  1.0000  1.0000

where each entry in a column of x has the same value and each entry in a row of y has the same value. Once these two arrays have been created, we can create an identically sized array containing values of the function with the single statement, say

>> z = x.^2 + y.^2
z =
10.0000  2.0000  2.0000  10.0000
  9.2500  1.2500  1.2500  9.2500
  9.0000  1.0000  1.0000  9.0000
  9.2500  1.2500  1.2500  9.2500
 10.0000  2.0000  2.0000  10.0000

(Don’t overlook the dot before the symbol ^ in this statement; we want element-by-element multiplication, not matrix multiplication.) Here, each entry in the array z is the value of the function \( z = x^2 + y^2 \) at the point whose x coordinate is at the corresponding position in the array x and whose y coordinate is at the corresponding position in the array y.

To illustrate both the use of meshgrid and the nature of its output, we above chose an example that is far too small to produce a useful display of any function. We would create a more useful display if the array providing input to the several graphing routines had more entries. To prepare for subsequent examples, then, we execute the statements

>> dx = 1.0/40.0;
>> [x,y] = meshgrid( 0.0:dx:1.0, 0.0:dx:1.0 );
>> z = sin( 2.0*pi*x ) .* sin( 3.0*pi*y );
>> dxi = 6.0*pi/49.0;
>> [xi,eta] = meshgrid( -3*pi:dxi:3*pi, -3*pi:dxi:3*pi );
>> I = ( (sin(xi)./xi).^2 ) .* ( (sin(eta)./eta).^2 );

to create more refined arrays z and I for the two examples of Eqs. (4.10) and (4.11). (The “funny” increment in the second example places 49 divisions—rather than 50—in the interval \(-3\pi \leq \xi, \eta \leq 3\pi\) and is chosen to avoid the appearance of the explicit value 0 in either xi or eta, thereby avoiding several divisions by zero in evaluating I.)

4.13.2 Surface Plots: The Function mesh

The simplest OCTAVE function for displaying two-dimensional arrays is surf, and the statements

>> mesh( z )
>> mesh( I )
4.13. GRAPHING SCALAR FUNCTIONS OF TWO VARIABLES

Figure 4.10: A wire mesh representation of the function $z$ in Eq. (4.10).

will use the default colormap to generate a colored, three-dimensional wire mesh graph conveying the altitudes in the two-dimensional arrays $z$ and $I$. The specification of additional properties, however, will create a more refined display. Executing the more elaborate statements\(^{57}\)

\[
\text{>> mesh}( x, y, z, 'edgecolor', 'black' )
\text{>> title( '[2,3] Mode', 'fontsize', 20 )}
\text{>> xlabel( 'x', 'fontsize', 16 )}
\text{>> ylabel( 'y', 'fontsize', 16 )}
\text{>> zlabel( 'z', 'fontsize', 16 )}
\text{>> set( gca, 'fontsize', 14 )}
\text{>> mesh}( xi, eta, I, 'edgecolor', 'black' )
\text{>> title( 'Diffraction (Square Aperture)', 'fontsize', 20 )}
\text{>> xlabel( '\xi', 'fontsize', 16 )}
\text{>> ylabel( '\eta', 'fontsize', 16 )}
\text{>> zlabel( 'I', 'fontsize', 16 )}
\text{>> set( gca, 'fontsize', 14 )}
\]

for example, will embellish the simple displays by labeling the axes and titling the graph with characters that are a bit larger than the default characters. Here, the first and second arguments to \text{mesh} (x or xi; y or eta) tell OCTAVE to use the elements of these arrays to scale the two axes, so the axes will be labeled to show a range of $0.0 \leq x, y \leq 1.0$ or $-3\pi < \xi, \eta < 3\pi$ (rounded to $-10.0 \leq \xi, \eta \leq 10.0$). The third argument provides the values of the function to be represented as a surface, and the specification of the property \text{edgecolor} overrides the coloring of the lines and makes them all black, facilitating printing on a monochrome printer. The more complete graphs produced by these last statements are shown in Figs. 4.10 and 4.11. Note that, in the perspective shown, the $x$ and $\xi$ axes run positively to the east northeast, the $y$ and $\eta$ axes run positively to the west northwest, and the $z$ and $I$ axes run positively to the north (vertical).\(^{58}\)

\(^{57}\)In the second of the statements involving \text{xlabel} and \text{ylabel}, we have invoked OCTAVE’s capability to insert Greek letters in a display. This feature is more fully described in Section 4.17.1.

\(^{58}\)Note that using the device \text{-deps} in the \text{print} command renders the black and white display on the screen in a PostScript file that uses red for positive values of $z$ and $\xi$ and green for negative values. To create a black and white
The `mesh` command produces a particular perspective on the surface by default. To change that perspective on the screen, simply pick up the surface graph with the left mouse button and drag the view to any other orientation. That rotated orientation, however, will not be transferred to a PostScript file created with `print -deps FileName.eps`. To rotate the on-screen display in a way that will be transferred to a PostScript file, execute the statement

\[
\text{view}(\text{azimuth, elevation}) \quad \text{or} \quad \text{view}([\text{azimuth, elevation}])
\]

where `azimuth` and `elevation` specify in degrees the azimuth and elevation of the point of view. (The statement `view(3)` will restore the default values -37.5 and 30, respectively for 3-D displays.) This statement can be executed at any time and will immediately affect the current `Figure` window. If, for example, we execute the command

```matlab
>> view(30, 40)
```

after each of the displays in Figs 4.10 and 4.11 has been created, the results will be as shown in Figs 4.12 and 4.13.

Examination of Figs. 4.11 and 4.13 suggests that the pattern might have weak peaks in areas outside the main central peak. Autoscaling of the vertical (`I`) axis, however, is influenced by the high central peak. To show the secondary peaks more clearly, we could make use of the command `axis` to specify a scale on the vertical axis that ignores the central peak. For example, the statement

```matlab
>> axis([-10.0 10.0 -10.0 10.0 0.0 0.15])
```

preserves the automatically determined ranges on the \(\xi\) and \(\eta\) axes but forces the \(I\) axis to be confined to the range \(0.0 \leq I \leq 0.15\) (and the central peak simply goes off scale). The resulting display, shown in Fig. 4.14, reveals the secondary peaks in the diffraction pattern and even hints at some of those not on the main axes.

PostScript file, use the device `-depsc`, which will preserve the “colors”, whether all black or something else, in the PostScript file.
4.13.3 Shaded Surfaces: The Function surf

An alternative presentation of a surface is provided by the command surf, which draws a wire mesh but then shades the surface by coloring each element defined by the mesh with a color appropriate to the value of the function at that element. Given the evaluation of the functions $z$ and $I$ as in Section 4.13.1, the statements

```
>> surf( x, y, z )
>> surf( xi, eta, I)
```

will produce shaded surface displays of the two surfaces using the default association of colors with values. If you wish to change the translation of values into colors different from those used by default, invoke the command colormap to select an alternative color map from the built-in options\(^\text{59}\) that will be displayed by the statement colormap('list'). In particular, executing the statement colormap('gray') will produce a gray scale on the screen, the statement colormap('white') will suppress the filling of each element in the surface plot altogether, leaving only a wire mesh diagram in which the lines are all black, as shown in Figs. 4.10 and 4.11, and the statement colormap('default') restores the default color map.

The command shading offers a additional control over the way the surface plot is drawn. Specifically,

- `shading( 'faceted' )`—the default—makes each patch the color appropriate to the value at that patch and displays the edges in black.
- `shading( 'flat' )` makes each patch the color appropriate to the value at that patch and suppresses display of the edges.
- `shading( 'interp' )` interpolates colors between patch vertices and suppresses the edges.

Thus, to produce the display in Fig. 4.15, which suppresses both the grid lines and selects a gray shading, execute the statements

```
% See the OCTAVE manuals for information about how to create a custom color map.
```
Figure 4.13: An alternative view of a wire mesh representation of the function $I$ in Eq. (4.11).

![Diffractive (Square Aperture)](image1)

Figure 4.14: A rescaled surface representation of the function $I$ in Eq. (4.11).

![Diffractive (Square Aperture)](image2)
Figure 4.15: A shaded surface representation of the function $z$ in Eq. (4.10). Curiously, in some versions of OCTAVE, some routes to a PostScript file for this display results in pieces of this display that are (properly) hidden in the on-screen PLOT window are shown when the PostScript rendition of the display is created.

>> surf( x, y, z, 'edgecolor', 'black' )
>> colormap( 'gray' )
>> shading( 'interp' )
>> title( '[2,3] Mode', 'fontsize', 20 )
>> xlabel( 'x', 'fontsize', 16 )
>> ylabel( 'y', 'fontsize', 16 )
>> zlabel( 'z', 'fontsize', 16 )
>> set( gca, 'fontsize', 14 )

### 4.13.4 Contour Plots: The Function contour

The function `contour` draws contour lines for a two-dimensional array. We can produce a meaningful contour map with the statements

```matlab
>> colormap( 'default' )
>> contour( z, 10 )
>> contour( x, y, z, 10 )
>> contour( x, y, z, [-0.8,-0.6,-0.4,-0.2,0.2,0.4,0.6,0.8] )
```

Here, the first statement makes sure the default colormap is in place; the second statement produces a contour plot of the function $z(x,y)$, using 10 contour lines and labeling the axes with the indices of the entries in the matrix $z$; and the third statement adjusts the display produced by the second statement by labeling the axes with the actual values of $x$ and $y$. In contrast to the automatic choice of the values of $z$ at which contour lines are drawn, the fourth statement stipulates explicitly the values to be used. With the default colormap, each contour line is colored to convey the value of $z$ at each line. Alternatively, the statements

```matlab
>> contour( x, y, z, 10, 'k' )
>> axis equal
```
CHAPTER 4. INTRODUCTION TO OCTAVE

Figure 4.16: Contour plot of the function $z$ in Eq. (4.10). This view shows the membrane from above. Contour lines at positive values of $z$ are solid; those at negative values of $z$ are dashed.

override the current colormap, displaying each line with the solid color black, and stipulate that equal values be represented by equal increments on both axes (which results in an undistorted view of the square membrane). (The second statement must follow the first.) Further, the statements

\[
\text{>> contour}( x, y, z, 10, 'k--' )
\]
\[
\text{>> axis equal}
\]

will produce a graph in which the contour lines are drawn dashed and the statements

\[
\text{>> contour}( x, y, z, [0.8, 0.6, 0.4, 0.2], 'k' )
\]
\[
\text{>> axis equal}
\]
\[
\text{>> hold}
\]
\[
\text{>> contour}( x, y, z, [-0.8, -0.6, -0.4, -0.2], 'k--')
\]
\[
\text{>> set( gca, 'fontsize', 16 )}
\]
\[
\text{>> hold off}
\]

will produce a graph in which contour lines having positive values of $z$ are solid and those having negative values of $z$ are dashed. The display resulting from these last four statements is shown in Fig. 4.16.

Even better, we could draw and label the positive and negative contour lines. To do so, we must exploit the capacity of contour to return both contour matrices ($\text{Cpos}$ and $\text{Cneg}$) and handles ($\text{Hpos}$ and $\text{Hneg}$) which are needed for the new command clabel. The statements

\[
\text{>> [Cpos, Hpos] = contour}( x, y, z, [0.8, 0.6, 0.4, 0.2], 'k' );
\]
\[
\text{>> axis equal}
\]
\[
\text{>> clabel( Cpos, Hpos )}
\]
\[
\text{>> hold on}
\]
\[
\text{>> [Cneg, Hneg] = contour}( x, y, z, [-0.8, -0.6, -0.4, -0.2], 'k--');
\]
\[
\text{>> clabel( Cneg, Hneg )}
\]
\[
\text{>> hold off}
\]

will produce the labeled display, though adding the labels on the contour lines yields a display that is a bit too busy for clarity.
4.13.5 The Functions meshc, surfc, and surfl

A more complex way to represent a two-dimensional array would invoke one of the functions meshc, surfc, and surfl. In a single three-dimensional box, meshc displays a mesh surface and, along the bottom of the box, a contour plot. The behavior of surfc is similar, though it displays a shaded surface rather than a mesh surface. In contrast, surfl creates much the same display as surf, except that it illuminates the surface with a light source. The statements

\[
\text{>> meshc}(x, y, z) \\
\text{>> meshc}(xi, eta, I) \\
\text{>> surfc}(x, y, z) \\
\text{>> surfc}(xi, eta, I) \\
\text{>> surfl}(x, y, z) \\
\text{>> surfl}(xi, eta, I)
\]

will produce these displays for each of our sample functions. The details for using meshc, surfc, and surfl are fully described in the OCTAVE manuals.

4.13.6 Functions of Two Variables in Polar Coordinates

OCTAVE also makes available the function pol2cart to facilitate plotting functions of two variables expressed in polar coordinates. This function accepts matrices defining the coordinates of a grid in polar coordinates and converts them into matrices representing the same points in Cartesian coordinates. For example, the statements

\[
\text{>> } \phi, r = \text{meshgrid}(0.0:10.0:360)*\pi/180, 0.0:0.2:5.0; \\
\text{>> } x, y = \text{pol2cart}(\phi, r);
\]

create arrays \(\phi\) and \(r\) conveying a grid of points in the intervals \(0 \leq \phi \leq 2\pi\) (in radians) and \(0.0 \leq r \leq 5.0\) and then convert those arrays into arrays giving the Cartesian coordinates of the same points. Then, the statements

\[
\text{>> } z = (1 - r^2) \cdot e^{-r} \cdot \cos(\phi); \\
\text{>> mesh}(x, y, z, 'edgecolor', 'black') \\
\text{>> set(gca, 'fontsize', 16) \\
\text{>> contour}(x, y, z, 20, 'linecolor', 'black') \\
\text{>> set(gca, 'fontsize', 16) \\
\text{>> axis square} \\
\text{>> surf}(x, y, z)
\]

will produce a mesh surface, a contour map, and a shaded surface depicting the function

\[
z(r, \theta) = (1 - r^2) e^{-r} \cos \theta \tag{4.12}
\]

In effect, the command pol2cart translates polar coordinates to Cartesian coordinates so that we can create polar plots by using the commands mesh, contour, and surf. The first two of these displays are shown in Fig. 4.17 and Fig. 4.18.

The purpose of this brief section is to draw attention to the existence of the command pol2cart. Details of its use can be found in the OCTAVE manuals.
4.14 Graphing Scalar Functions of Three Variables

Scalar functions of three variables also occur regularly in physics. Displaying them graphically, however, is complicated because there are four quantities (three independent variables; one dependent variable) to be conveyed. The task involves developing ways to represent four-dimensions in three and then, even more, to project those three onto the two-dimensional screen of a computer workstation. However that is accomplished, the input to the process will have to be a three-dimensional matrix, each of whose entries gives the value of the function at the point \((x_{ijk}, y_{ijk}, z_{ijk})\) in a regular, three-dimensional grid of points spanning the region of space within which we want to examine the function.

For the sake of a more specific example, we choose the normalized probability density \(p(x, y, z)\) for the electron in the \((n, l, m) = (3, 2, 0)\) state in the hydrogen atom. This probability density is
given as a function of Cartesian coordinates \((x, y, z)\) with the nucleus located at the origin by

\[
p(x, y, z) = \frac{1}{2\pi(27)^{3/2}} \rho^4 e^{-2\rho/3} \left( \frac{3z^2}{\rho^2} - 1 \right)^2 = \frac{1}{2\pi(27)^{3/2}} e^{-2\rho/3} (9z^4 - 6z^2 \rho^2 + \rho^4)
\]

(4.13)

where the coordinates are all measured in units of the Bohr radius and

\[
\rho = \sqrt{x^2 + y^2 + z^2}
\]

(4.14)

To be explicit, we determine values of \(p(x, y, z)\) over the region \(-10.0 \leq x, y, z \leq 10.0\), dividing each axis into 32 segments, which will entail evaluating \(p(x, y, z)\) over a grid containing 33 values of \(x\), 33 values of \(y\), and 33 values of \(z\)—a total of \(33 \times 33 \times 33 = 35937\) points.

Using OCTAVE to generate a three-dimensional array containing values of this probability density at a grid of points in the desired region is straightforward, since the command `meshgrid` can accept not only two ranges (to produce two two-dimensional arrays) but also three ranges (to produce three three-dimensional arrays). An array containing values of the probability density in Eq. (4.13) is created using the statements

\[
\begin{align*}
&\text{Calculate increment.} \\
&\text{Calculate grid.} \\
&\text{Calculate premultiplier.} \\
&\text{Calculate } z^2. \\
&\text{Calculate } \rho^2. \\
&\text{Calculate } \rho. \\
&\text{Calculate } p(x, y, z).
\end{align*}
\]

\[
\begin{align*}
\text{inc} & = 20.0/32.0; \\
[x, y, z] & = \text{meshgrid}( -10.0 : \text{inc} : 10.0, \ldots) \\
pfactor & = 1.0/(2.0*\pi*27.0^{3/2}); \\
zs & = z.^2; \\
rhos & = x.^2 + y.^2 + zs; \\
rho & = \text{sqrt}(rhos); \\
p & = pfactor * \exp(-2.0*rho/3.0) .* \ldots \\
& (9.0*zs.^2 - 6.0*zs.*rhos + rhos.^2);
\end{align*}
\]

The first statement establishes the coordinate increment. The second statement—which appears distributed over three physical lines—produces a grid of values for the evaluation of the scalar field. The remaining statements evaluate several intermediate quantities and then evaluate \(p\), which is a three-dimensional array that will be the input for the production of a variety of displays.

### 4.14.1 Reduction to Two-Dimensional Displays

One way to show the characteristics of a function of three variables is to display two-dimensional subsets of the three-dimensional data, that is, to examine the behavior of the function in planes that intersect the volume. The simplest planes to extract are those for which \(z\) has a fixed value, since the expression \(p(:,:,5)\), for example, will extract a two-dimensional array containing the values of \(p(x, y, z)\) for which the third coordinate is five (or—better—whatever value of \(z\) corresponds to the integer 5 in the original scaling, \(z = z_0 + 5 \times dz = -10.0 + 5 \times 20.0/32.0 = -6.875\)). Thus, the statements

\[
\begin{align*}
&\text{mesh( } x(:,:,5), y(:,:,5), p(:,:,5) ) \\
&\text{contour( } x(:,:,5), y(:,:,5), p(:,:,5) ) \\
&\text{surf( } x(:,:,5), y(:,:,5), p(:,:,5) )
\end{align*}
\]

will produce mesh surface, contour, and shaded surface representations of the probability density in the plane parallel to the \(xy\) plane at the specified value of \(z\).

Extracting data in a plane perpendicular to the \(x\) or \(y\) axes is more complicated, since the expression \(p(15,:,:)\) will not achieve the desired end. Instead, we must be more deliberate. The statements

\[
\begin{align*}
&\text{We must extract two-dimensional arrays from all three arguments, since these statements will not behave properly with three-dimensional arrays in any position.}
\end{align*}
\]
for example, will extract data in a plane parallel to the $yz$ plane and generate first a mesh surface and then a contour map of those data. Note that we have here recognized that the arguments giving the scale for `mesh` and `contour` can be vectors (rather than higher-dimensional arrays) containing the coordinates of the points along the axes. The figures produced by these commands are presented in Fig. 4.19 and Fig. 4.20.

While quick, the method of the previous paragraph is restricted to planes in which one has data, to planes perpendicular to one or another of the coordinate axes, and to a small portion of the data. OCTAVE’s command `slice` is more powerful, since it allows easy positioning of one or more planes that “slice” through the three-dimensional volume and also permits oblique slices. To draw slices parallel to the coordinate planes, the general syntax of a statement invoking `slice` is

\[
\text{slice}(x, y, z, V, S_x, S_y, S_z);
\]

where the three-dimensional arrays $x$, $y$, and $z$ give the coordinates of the points at which the function $f(x,y,z)$ is evaluated, the three-dimensional array $V$ contains the values of the function at those points, and $S_x$, $S_y$, and $S_z$ (which may be single values or vectors—or empty []) give the values of the coordinates at which slices are to be drawn. Thus, for example, the statements

\[
\text{Create a floating point array.}
\]
Figure 4.20: Contour map of the function $p$ in Eq. (4.13) in the plane $x = x_0 + 15 \times dx = -10.0 + 15 \times 20.0/32.0 = -0.625$.

```
>> axis square
>> slice( x, y, z, p, 0.0, 0.0, 0.0 );
>> shading interp
>> set( gca, 'visible', 'off' )
```

set the axes appropriately; display three slices, one parallel to the $yz$ plane at $x = 0$, a second parallel to the $zx$ plane at $y = 0$, and a third parallel to the $xy$ plane at $z = 0$; smooth the shading; and exploit the axes property `visible` to turn off the axes. Alternatively, the statements

```
>> axis square
>> slice( x, y, z, p, [-7.5, 0.0, 7.5], 0.0, 0.0 );
>> shading interp
>> set( gca, 'Visible', 'off' )
>> colorbar
```

produce three slices parallel to the $yz$ plane at $x = -7.5$, 0.0, and 7.5, one parallel to the $zx$ plane at $y = 0$, and one parallel to the $xy$ plane at $z = 0$ and add a colorbar to facilitate interpreting the color code.

Placing an oblique slice on the display involves a bit more work. We must first create arrays containing the coordinates of points on the surface defining the slice by drawing a simple plane (say the plane $z = 0$), rotate it to the desired orientation, and then read the desired coordinates from the rotated display. Appropriate statements are

```
>> [xx,yy] = meshgrid(-10.0:0.625:10.0, ...
   -10.0:0.625:10.0);
>> hdl = surf( xx, yy, zeros(33,33) );
>> axis( [-10.0, 10.0, -10.0, 10.0, ...
   -10.0, 10.0] );
>> rotate(hdl, [0,1,0], -90 );
>> rotate(hdl, [0,0,1], 45 );
>> XI = get( hdl, 'XData' );
>> YI = get( hdl, 'YData' );
>> ZI = get( hdl, 'ZData' );
```

Create coordinates for drawing $xy$ plane.
Draw surface, saving handle.
Scale axes appropriately.
Rotate $-90^\circ$ about $y$ axis.
Rotate $45^\circ$ about $z$ axis.
Get data defining rotated surface.
Here we have used the handle `hdl` to the surface several times, first in the command `rotate` (which rotates the identified object by the angle in the third argument about the axis whose direction is given by the (vector) second argument) and then in the command `get` (which reads coordinates of points defining the rotated surface into an appropriately structured array).

With the arrays `XI`, `YI`, and `ZI` now in hand, we can generate the desired display with the additional statements

```matlab
>> close all
>> slice( x, y, z, p, XI, YI, ZI )
>> hold on
>> slice( x, y, z, p, 0.0, 0.0, []);
>> shading interp
>> colorbar
>> axis square
>> set( gca, 'Visible', 'off' )
>> colormap( 'gray' )
>> hold off
```

Remove `Figure` window.
Draw oblique slice.
Suppress erase.
Add “ordinary” slices, suppressing any parallel to the `xy` plane.
Set shading.
Add colorbar.
Equalize axes.
Turn off display of axes.
Set gray scale.

Figure 4.21 shows the display produced by these statements, though the black and white rendition pales by comparison with the color rendition on the screen.

### 4.14.2 Isosurfaces

Another method of viewing a three-dimensional scalar field is to look at isosurfaces, or sets of locations at which the function has a constant value. In our present example, we locate and display a surface of constant probability density with the statements
Figure 4.22: An isosurface of the probability density $p$ in Eq. (4.13) using `isosurface` but suppressing shading.

```matlab
>> close all
>> poly = isosurface( x, y, z, p, 1.2e-4 );
>> hpatch = patch(poly);
>> view( 3 )
>> set( gca, 'fontsize', 16 )
```

Remove `Figure` window.
Find isosurface.
Display isosurface, saving handle.
Set to default viewing position.
Adjust fontsize for labels on axes

The command `isosurface` searches the array $p$ to generate in `poly` information about the faces and vertices composing the isosurface at which $p(x, y, z)$ has the specified value $1.2 \times 10^{-4}$. (Other values could be used.) The command `patch` draws the isosurface in a `Figure` window, though it doesn’t generate a suitable perspective in that window. Thus, we need the statement `view(3)` to establish an appropriate viewing angle—here the default—on the surface. The resulting output is not shown because, unfortunately, the shape of this surface is not very transparent in the figure produced in this way. Replacing the last three of the above statements with the statements

```matlab
>> hpatch = patch(poly, 'edgecolor', 'black', 'facecolor', 'none');
>> view( 3 )
>> set( gca, 'fontsize', 16 )
```

we can suppress the shading of the surface to produce the three-dimensional wire diagram in Fig. 4.22. The shape of the surface is rather more discernible in this second figure. Note that, behind the scenes, the operations producing these displays involve much calculation; be patient, since it may take a while for the calculations to be completed.

---

61 We should, of course, be familiar with the range of the data so that the chosen value actually occurs within the data. Note the functions `max` and `min`, which will extract the largest and smallest values from a vector. To find the largest value of $p$, for example, we might execute the statement `max(max(max(p)))`. 
4.15 Graphing Vector Fields

Graphical display of vector functions of two or three variables is more complicated than display of scalar functions, but OCTAVE has at least one function designed to produce such displays. To illustrate the general procedures briefly and without much detail, let us consider the electromagnetic field in a transverse electric (TE) wave in the rectangular waveguide shown in Fig. 4.23 when the electric field has only a z component, the magnetic field has only x and y components, and none of these components depends on z. In this case, the only non-zero components of the electric field $\mathbf{E}$ and the magnetic field $\mathbf{H}$ are given at time $t = 0$ by

\begin{align*}
E_z &= \cos \kappa_x b \pi \sin n \pi \gamma \\
H_x &= \sin \kappa_x b \pi \cos n \pi \gamma \\
H_y &= -\kappa_x b \cos \kappa_x b \pi \sin n \pi \gamma
\end{align*}

(4.15)

where the fields are measured in units in which the amplitude of $E_z$ is one. Further, $b$ is the y-dimension of the guide, $n$ is a positive integer, $\pi = x/b$ and $\gamma = y/b$ are dimensionless coordinates within the guide, and

\[
\left( \frac{\kappa_x b}{n \pi} \right)^2 = \left( \frac{\omega b}{n \pi c} \right)^2 - 1
\]

(4.16)

where $c$ is the speed of light and $\omega$ is the frequency of the wave. The field depends on two parameters, $n$ and $\kappa_x b$, where $\kappa_x b$ is determined from Eq. (4.16) by $\omega$, $b$, and $c$. Since $\kappa_x b$ must be real if the wave is to propagate down the guide, we must require that $(\omega b/n \pi c)^2 > 1$ or that $\omega > n \pi c/b$. For the sake of a specific example, we choose $b = 1$ and $n = 2$, and then we choose $\omega$ so that $\kappa_x b/2\pi$ turns out to have the value 1. With these choices, the fields are given by

\begin{align*}
E_z &= \cos 2\pi \pi \sin 2\pi \gamma \\
H_x &= \sin 2\pi \pi \cos 2\pi \gamma \\
H_y &= -\cos 2\pi \pi \sin 2\pi \gamma
\end{align*}

(4.17)

where $\pi$—the coordinate along the guide can range over any values—we choose $0 \leq \pi \leq 1$—but, to be inside the guide, $\gamma$ is confined to the region $0 \leq \gamma \leq 1$.

\footnote{For a discussion of the electromagnetic fields in wave guides, see The Theory of the Electromagnetic Field by David M. Cook (Prentice-Hall, Englewood Cliffs, NJ, 1975) or Introduction to Electrodynamics, Third Edition, by David J. Griffiths (Prentice-Hall, Upper Saddle River, NJ, 1999). The first of these books, out of print since the early 1990’s, was available for awhile after January, 2003, in a Dover reprint, but that reprint has since not been kept in print.}
Each component of these fields can now be represented by a two-dimensional array. The \( \mathbf{H} \) field, which has two non-zero components, then is translated into two such arrays; the \( \mathbf{E} \) field, which has only one non-zero component, requires only one such array. These arrays are readily created with the OCTAVE statements\(^63\)

\[
\begin{align*}
&\text{>> } [x, y] = \text{meshgrid}(0.0:0.04:1.0, 0.0:0.04:1.0); \\
&\text{>> } xp = 2.0\pi x; \quad yp = 2.0\pi y; \\
&\text{>> } Ez = \cos(xp) .* \sin(yp); \\
&\text{>> } Hx = \sin(xp) .* \cos(yp); \\
&\text{>> } Hy = -\cos(xp) .* \sin(yp); \\
\end{align*}
\]

### 4.15.1 The Function `quiver`

The OCTAVE function `quiver` produces a two-dimensional vector field plot. At each grid point, `quiver` draws an arrow which conveys both the direction and the magnitude of the field at that point. The calling statement for this function is

\[
\text{quiver}(x, y, u, v)
\]

where \( x \) and \( y \) are either vectors containing the actual coordinates of the grid points on the two axes or two-dimensional arrays containing the coordinates of all points in the grid at which vectors are to be drawn, and \( u \) and \( v \) are the \( x \) and \( y \) components of the two-dimensional field, respectively, at the grid points conveyed by \( x \) and \( y \). Thus, for example, the two components of \( \mathbf{H} \) in OCTAVE’s memory can be displayed in this form by executing the statements

\[
\begin{align*}
&\text{>> quiver}(x, y, Hx, Hy, 'k') \\
&\text{>> axis( [0 1 0 1] )} \\
&\text{>> axis square} \\
&\text{>> xlabel('x', 'fontsize', 16) } \\
&\text{>> ylabel('y', 'fontsize', 16)} \\
&\text{>> set( gca, 'fontsize', 16) }
\end{align*}
\]

The resulting display is shown in Fig. 4.24.

### 4.15.2 More Elaborate Displays

Numerous additional displays can be created. We illustrate the possibilities with a plot in which the \( x \) and \( y \) components of the magnetic field displayed with `quiver` are superimposed on a contour map of the \( z \) component of the electric field. The statements

\[
\begin{align*}
&\text{>> hold on} \\
&\text{>> contour}(x, y, Ez, [0.1,0.3,0.5,0.7,0.9], 'k') \\
&\text{>> contour}(x, y, Ez, [-0.9,-0.7,-0.5,-0.3,-0.1], 'k--') \\
&\text{>> hold off}
\end{align*}
\]

will add contour lines to Fig. 4.24, producing the display shown in Fig. 4.25. Here, we have used arguments to `contour` to specify that contours for positive values should be drawn with a solid line (the default linestyle) while contours for negative values should be drawn with a dash-dot line (linestyle ‘-.’). Thus, the electric field is coming out of the page in regions where its component is indicated with a solid line and going into the page in regions where its component is indicated with a dash-dot line.

\(^{63}\)From here on in this section, we drop the overbars.
Figure 4.24: Magnetic field of Eq. (4.17) at $t = 0$ in a rectangular wave guide using quiver. The illustrated structure, of course, propagates along the guide toward positive $x$ as time unfolds.

Figure 4.25: Plot showing both magnetic and electric fields of Eq. (4.17) at $t = 0$ in wave guide. The magnetic field lines are closed curves in the plane of the paper. The electric field is directed perpendicular to the paper and has at each point a component whose magnitude is conveyed by the contour lines. The illustrated structure, of course, propagates along the guide toward positive $x$ as time unfolds.
4.15.3 Three-Dimensional Vector Fields

When a vector field has non-zero components in more than two coordinate directions, a picture of the field is much harder to draw and even more difficult to interpret without the aid of a tool that will allow easy viewing of the picture from a variety of viewing angles. One can, of course, use `quiver` to draw pictures of the projections of the field in any given plane into that plane. In addition, OCTAVE provides the function `quiver3` (see the OCTAVE manuals) for drawing a picture similar to those produced by `quiver`, but drawing short vectors at points in three-dimensional space. The primary input to `quiver3` is a sextet of three-dimensional arrays giving the $x$, $y$, and $z$ coordinates of the points at which vectors are to be drawn and the $x$, $y$, and $z$ components of the vector field at those points. The exploration of these capabilities is left to the exercises. Remember that the mouse can be used to rotate 3D-graphs produced by OCTAVE in the `Figure` window.

4.16 Animation

Along with all of the image processing and graphing capabilities, more recent versions of OCTAVE can also display a sequence of images in rapid succession to produce an animated display.\(^{64}\) In essence, one creates each image in turn in an appropriately sized window, reads each window as a bit-map into an element of an appropriately sized array, and then displays each of those images in turn.

To illustrate this process, we use a data file containing 41 $33 \times 33$ arrays, each of which conveys the displacement of an L-shaped membrane at a particular moment in time.\(^{65}\) The statements creating a three-dimensional byte array containing the 41 images in the animation are

\[
\begin{align*}
\text{	exttt{>> id = fopen('ldrum.dat', 'r');}} & \quad \text{Open data file.} \\
\text{	exttt{>> x = [ 0.0 : 1.0/32.0 : 1.0 ];}} & \quad \text{Create vector of } x \text{ values.} \\
\text{	exttt{>> for j=1:41}} & \quad \text{For each of 41 frames:} \\
& \quad \text{	exttt{disp = fscanf( id, '%f', [33,33] );}} \quad \text{Read array from file.} \\
& \quad \text{	exttt{mesh( x, x, disp', 'edgecolor', 'black' )}} \quad \text{Create display with transposed array.}\(^{66}\) \\
& \quad \text{	exttt{axis( [ 0.0 1.0 0.0 1.0 -0.25 0.25 ] )}} \quad \text{Standardize axes.} \\
& \quad \text{	exttt{M(j) = getframe;}} \quad \text{Read display as bit map.} \\
\text{	exttt{end}} & \quad \text{Close file.} \\
\text{	exttt{>> status = fclose( id );}} & \quad \text{Close file.} \\
\end{align*}
\]

At this point, the array $M$, whose class (struct) is new, contains all of the images stored as bit maps so they can be displayed one after the other in quick succession.\(^{67}\)

Several options for displaying the images can be invoked. All use the command `movie` but with different arguments. The statement

\[
\text{	exttt{>> movie( M );}}
\]

displays each image in the array $M$ one after another from beginning to end and then stops. The statement

\[\text{\texttt{\textsuperscript{64} Necessary routines for easy animation were not present in Version 4 of OCTAVE. They have been added in Version 5.2 and may have been included in some versions before Version 5.2.}}\]
\[\text{\texttt{\textsuperscript{65} The file was created by using \texttt{lsode}—the Livermore solver for ordinary differential equations—to effect a numerical solution of the wave equation for the membrane.}}\]
\[\text{\texttt{\textsuperscript{66} Here, we have used the shorthand symbol \texttt{'} to transpose \texttt{disp} rather than the longer function \texttt{transpose}. Strictly, the symbol \texttt{'} effects the evaluation of the adjoint—the transposed complex conjugate—of the matrix. When the elements are all real, however, the adjoint and the transpose are the same. If we wanted to transpose a matrix of complex elements \texttt{without complex conjugation, we would use \texttt{transpose} or the special symbol \texttt{'}.'}.}}\]
\[\text{\texttt{\textsuperscript{67} Actually, each element of \texttt{M} is itself structured, having two components \texttt{M(j).cdata}, which contains the image, and \texttt{M(j).colormap}, which has no function in OCTAVE and is included for compatibility with MATLAB.}}\]
CHAPTER 4. INTRODUCTION TO OCTAVE

```octave
>> movie( M, 15 );
```

exploits an optional (integer) argument to effect a 15-fold display of the full sequence of images, and the statement

```octave
>> movie( M, [ 15 1 3 5 7 9 ] );
```

will display images 1, 3, 5, 7, and 9 in succession, 15 times over. As a special case, the statement

```octave
>> movie( M, [ 1 23 ] );
```

for example, will display image 23 once. Since that image is then left in the Figure window, this version of the command allows display of a single image for close examination. Finally, the statement

```octave
>> movie( M, [15 1 3 5 7 9], 2 );
```

exploits a further argument that controls the speed of the display—here set to 2 frames per second. As with all OCTAVE commands, we can use (CONTROL-C) to interrupt a movie before it is finished, though the OCTAVE Command window must be the active window for this action to succeed.

The sequence of statements described in the previous paragraph ultimately displays a succession of images, each of which is separately defined in the input data. An alternative procedure is better suited to the task of creating an animation that can be generated by explicit transformation of a single image in some simple way. In this case, only the data defining the initial image need be supplied. OCTAVE can then generate all remaining images from that initial image. For example, generating an animated display showing the simple harmonic motion of one of the normal modes of a square drumhead is especially easy if we have data on its initial shape. We simply create a succession of images, each generated from the initial image by multiplying all values defining that image by \( \cos 2\pi t \), where \( t \) is stepped in equal increments from 0 to 1 so the multiplier runs over one period. For example, to display the sinusoidal motion of a square drumhead whose initial shape is given by

\[
 u(x,y,0) = \sin 2\pi x \sin 3\pi y, \quad 0 \leq x,y \leq 1
\]

we might start with the statements

```octave
>> [ x, y ] = meshgrid( 0.0:0.04:1.0, ... 
    0.0:0.04:1.0 );
>> u = sin(2*pi*x) .* sin(3*pi*y);
>> t = [ 0.0 : 0.025 : 1.0 ];
>> for j = 1:41
    disp = u*cos(2*pi*t(j));
    mesh( x, y, disp, 'EdgeColor', 'black' );
    axis( [ 0.0 1.0 0.0 1.0 -1.0 1.0 ] );
    M(j)=getframe;
end
```

At this point, the necessary images are stored in the (structured) array \( M \). From here, the animated display itself can be created with the command `movie` as described in the previous paragraph. We simply have generated the necessary input data by calculating it from the initial shape rather than by reading it from a file.

---

\(^{68}\)If the integer is positive, the movie is played forward the specified number of times; if the integer is negative, the sequence of frames is played alternately forward and backward until the cycle—once forward and once backward—has been repeated \( \text{integer} \) times.

\(^{69}\)Whether we stop one image short of \( t = 1 \) or include the image for \( t = 1 \) depends on how the animation is to be shown. If the display moves forward from the first to the last image and then moves backward from the last to the first, the image for \( t = 1 \) should be included. If the display moves forward from the first to the last image and then jumps abruptly back to the first image, the image for \( t = 1 \) should be omitted. We here assume the first mode of display.
4.17 Advanced Graphing Features

In this section, we describe some of OCTAVE’s more sophisticated capabilities. Because these routines are more involved than those previously described, it would be best to master the basic graphics routines before moving on to these examples.

4.17.1 Fonts

Among the most useful of OCTAVE’s auxiliary features is a capability to choose characters from a wide variety of standard fonts (including Roman, Italic, Greek), to display bold characters and common mathematical symbols, and to position characters as subscripts and superscripts. The examples in this subsection are intended to introduce some of OCTAVE’s capabilities but are in no way a substitute for careful reading of the OCTAVE manuals.

At base, the appearance of characters created by those commands that place characters on the screen are determined by several properties, including:

- **color**, which can be specified by predefined words like red, yellow, green, blue, magenta, cyan, black, and white or, alternatively, by a three-component vector whose components range between 0 and 1 and give, in order, the proportion of red ([1 0 0]), green ([0 1 0]), and blue ([0 0 1]) in the output.
- **fontangle**, which is normal by default but can be changed to italic or oblique.
- **fontname**, which can assume any of the values AvantGarde, Bookman, Courier, Helvetica, Helvetica-Narrow, NewCenturySchlbk, Palatino, Symbol, Times-Roman, ZapfChancery, and ZapfDingbats. OCTAVE substitutes Courier for any name not in this list.
- **fontsize**, which is expressed in fontunits (points by default).
- **fontunits**, which is points by default but can be changed to inches, centimeters, normalized, or pixels.
- **fontweight**, which is normal by default but can be changed to light, demi, or bold.
- **interpreter**, which is tex by default but can be changed to none. The value latex is planned but not yet implemented. When the tex option is active, a fair number of \TeX and \LaTeX commands can be included in character strings. Note that all toolkits recognize these \TeX commands for display on screen but proper interpretation in PostScript and PDF files produced with the print command is not guaranteed (and may vary with the version of OCTAVE in use).\(^70\) As with \TeX and \LaTeX, the characters \{ and \} can be used to limit the scope of a command. In this context, strings need to be presented in single quotation marks.

These properties are set in the axes object, and can be changed by appropriate specifications in the command axes, e.g., in the statement

```plaintext
axes( 'fontname', 'palatino', 'fontsize', 16, 'xlim', [0.0 10.0], ... 'ylim', [-1.0 1.0] )
hold on
```

which will create the axes and label them and then prevent them from being overridden in a subsequent plot command. Alternatively, after axes have been created by execution of some other graphics command, e.g., plot, these properties can be set by analogous specifications with the command set( gca, ...), e.g.,

\(^70\)See the OCTAVE manuals for your version of OCTAVE for details.
set( gca, 'fontname', 'palatino', 'fontsize', 16, 'xlim', ...
    [0.0 10.0], 'ylim', [-1.0 1.0] )

Through inheritance, the specifications in the previous paragraph affect all characters in all children of the associated axes object. Some of those children, however, have their own copies of the specified properties. Statements like

\[ hdl = \text{get}( \text{gca}, \text{'}title\text{'} ) \]
\[ \text{set}( \text{hdl}, \text{'}fontname\text{'} , \text{'}symbol\text{'} , \ldots ) \]

report the handle for the text object containing—here—the title in the current axes object and then set the indicated properties of that object away from those inherited from the parent axes object.

The methods for changing fonts described in the previous two paragraphs affect the font for an \textit{entire} object (title, axis label, \ldots). When we wish to change the font for a few characters within a string that is otherwise to be displayed in the default font, we must stipulate the font changes with commands embedded in quoted strings. For example, the (illustrative but mildly silly) statement

\texttt{>> title( 'This \{\text{\textit{fontsize}\{12\}\text{alpha}\text{beta}}\} is a Title', 'fontsize', 20 )}

will produce a title in which the words ‘This’ and ‘a Title’ will be presented in the default font at 20 pt while—in a \LaTeX-style notation—the Greek characters $\alpha\beta$ will be presented in 12 pt characters. Here, braces play two roles: (1) they enclose the arguments of the command \texttt{\text{\textit{fontsize}}} and (2) they limit the scope of the change in font.\footnote{Those familiar with \LaTeX will identify this role with a similar role played by braces in that text-processing system.} \footnote{Note that proper interpretation of the \LaTeX-style specifications will not occur if the string is enclosed in \textit{double} quotation marks unless the backslashes are also doubled.}

Several other Greek letters and mathematical symbols are available with commands that mimic those used in \LaTeX. All \textit{lower case} Greek letters\footnote{Omicron is produced with \texttt{\textbackslash{o}}, not \texttt{\textbackslash{omicron}}.} (including \texttt{\textbackslash{vatheta}, \textbackslash{varpi}, \textbackslash{varsigma}, and \textbackslash{varphi}}) can be inserted with a command like \texttt{\textit{alpha}} (letter name in lower case preceded by a backslash); all \textit{upper case} Greek letters except those that look like upper case Roman letters (e.g., A, B, and H) can be inserted with a command like \texttt{\textit{Gamma}} (letter name with initial letter in upper case preceded by a backslash). Some of the other available symbols are listed in Table 4.6. Thus, for example, the statement

\texttt{title('\alpha\leq\sqrt{\beta^2+\gamma^2}', 'fontsize', 20 )}

will produce the title $\alpha \leq \sqrt{\beta^2 + \gamma^2}$. We have here also illustrated that superscripts can be produced with the character \texttt{^}; subscripts are created with the character \texttt{_}. Should either of these characters be literally needed in the output, the quoted text must include \texttt{\_} or \texttt{\^}.

Two further points: (1) While in-string selection of \textit{fontname} and \textit{fontsize} can be accomplished with the commands \texttt{\textit{fontname}} and \texttt{\textit{fontsize}}, parallel in-string specification of \textit{fontweight} (normal, bold, demi, light) and \textit{fontangle} (normal, italic, oblique) is not available. Instead, we must use one of \texttt{\textbf{bf}, \textit{sl}, \textit{it}, and \textit{rm}} to produce bold, slanted, italic, and Roman (normal) styles, respectively, and—to limit the scope of the indicated change—we enclose the characters in braces, e.g., \texttt{{\textbf{bf} BoldWord(s)}}. (2) The characters \texttt{\_}, \texttt{^}, \texttt{\}, \texttt{\}, and \texttt{\}} have been given special meanings; should we actually wish to display these characters in a title or label, we precede each in the quoted string with a backslash, i.e., we produce these characters by including \texttt{\_}, \texttt{\^}, \texttt{\}, \texttt{\}, and \texttt{\}}, respectively, in the quoted string.

In Section 4.13.2, we have already seen one use of this capability. We present one more example. Let us embellish the graph of the magnetic field first presented in Fig. 4.5 by (1) changing the font

\footnote{Omicron is produced with \texttt{\textbackslash{o}}, not \texttt{\textbackslash{omicron}}.}
4.17. ADVANCED GRAPHING FEATURES

Table 4.6: Some codes for use in character strings in OCTAVE. Codes in the first, third, and fifth columns identify strings that can be included to produce the symbols in the second, fourth, and sixth columns. A full listing of available codes is at http://www.gnu.org/software/octave/doc/v4.0.3/Use-of-the-interpreter-Property.html.

<table>
<thead>
<tr>
<th>Code</th>
<th>Symbol</th>
<th>Code</th>
<th>Symbol</th>
<th>Code</th>
<th>Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>\aleph</td>
<td>\N</td>
<td>\pm</td>
<td>\pm</td>
<td>\leq</td>
<td>\leq</td>
</tr>
<tr>
<td>\int</td>
<td>\int</td>
<td>\infty</td>
<td>\infty</td>
<td>\partial</td>
<td>\partial</td>
</tr>
<tr>
<td>\perp</td>
<td>\perp</td>
<td>\leftarrow</td>
<td>\leftarrow</td>
<td>\bullet</td>
<td>\bullet</td>
</tr>
<tr>
<td>\rangle</td>
<td>\rangle</td>
<td>\leftarrow</td>
<td>\leftarrow</td>
<td>\div</td>
<td>\div</td>
</tr>
<tr>
<td>\approx</td>
<td>\approx</td>
<td>\uparrow</td>
<td>\uparrow</td>
<td>\neg</td>
<td>\neg</td>
</tr>
<tr>
<td>\cdot</td>
<td>\cdot</td>
<td>\downarrow</td>
<td>\downarrow</td>
<td>\nabla</td>
<td>\nabla</td>
</tr>
<tr>
<td>\times</td>
<td>\times</td>
<td>\circ</td>
<td>\circ</td>
<td>\ldots</td>
<td>\ldots</td>
</tr>
<tr>
<td>\surd</td>
<td>\surd</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>\ldots</td>
<td>\ldots</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

used for the various titles and (2) replacing the title on the y axis with the equation itself. The statements

```matlab
>> dz = 8.0/100.0;
>> z = [-4.0 : dz : 4.0 ];
>> B = (1.0 + z.^2).^(-1.5);
>> plot( z, B, 'color', 'black', 'linewidth', 4 )
>> grid on
>> title('Magnetic Field on \{itz\} Axis', 'FontSize', 20 )
>> xlabel('Dimensionless Position, \{itz\}', 'FontSize', 16 )
>> ylabel( '\{it B\}({\it z} )=1/( 1+{\it z}^2 )^{3/2}', 'FontSize', 16 )
```

will produce the revised graph of the magnetic field shown in Fig. 4.26.

4.17.2 Space Curves

Sometimes it is desirable to view the trajectory of a particle in three-dimensional space. OCTAVE has the ability of taking vectors of the \(x, y,\) and \(z\) coordinates of points on a three-dimensional path, projecting these points onto the two-dimensional screen, and then connecting consecutive points with lines. To illustrate this feature, consider the equations

\[
\begin{align*}
    x &= \cos t \\
    y &= \sin t \\
    z &= \alpha t
\end{align*}
\] (4.18)

where \(\alpha\) is a constant, describing the trajectory of a charged particle moving in a constant magnetic field directed along the \(z\) axis. The starting point is, of course, to evaluate \(x, y,\) and \(z\). Then, we invoke the function `plot3` (not to be confused with the plain vanilla function `plot`) to plot the graph. The statements

```matlab
>> t = 30.0*[ 0.0 : 0.005 : 1.0 ];
>> alpha = 0.5;
>> x = cos(t); y = sin(t); z = alpha*t;
>> plot3( x, y, z, 'color', 'black', 'linewidth', 4 )
>> set( gca, 'fontsize', 16 )
```

Note particularly the absence of a space following \(\\\text{lt}\) in these statements. Evidently, OCTAVE does not require a space to separate the LATEX command from the following text. Note also that the extra space before each \(z\) in Fig. 4.26 is not present in the on-screen display.
Figure 4.26: A more elegant graph of the $z$ axis magnetic field of a current carrying loop.

$$B(z) = \frac{1}{1 + z^2}^{3/2}$$

will produce the graph in Fig. 4.27.

### 4.17.3 Using Multiple Windows

In the simple circumstances so far introduced, we have accepted OCTAVE’s willingness to create a *Figure* window (Figure 1) when the command `plot` is invoked or to use an existing window if it has already been created by a previous command. We can manipulate additional windows with statements like$^{75}$

```matlab
hd1 = figure
hd1 = figure( 'position', ... [400,200,500,300] )
hd1 = figure( 'menubar', 'none' )
hd1 = figure( 'name', 'New Window' )
figure(hd1)
```

Create window with available free handle, returning handle in `hd1`.

Create window with its lower left corner located 400 pixels to the right of and 200 pixels above the lower left corner of the screen and with size 500 pixels wide and 300 pixels high.

Create window *without* menubar.

Create window with specified title at top.

Make window `hd1`; create it and make it current if it doesn’t exist.

$^{75}$Normally, of course, each of these statements would be terminated with a semicolon to suppress explicit output.
Figure 4.27: Space curve of a charged particle moving in a constant magnetic field along the $z$ axis.

```matlab
hdl = gcf                    % Get handle of current window.
clf                         % Clear current window but leave window in place.
get(hdl)                    % List current values of properties of window hdl.
set(hdl)                    % List properties and possible values for window hdl.
close(hdl)                  % Delete current window hdl.
```

Note that execution of the statement `figure(2)`, say, when the window with handle 2 already exists does not change the contents of that window; it merely makes it the current window.

Plot commands always direct their output to the current plot window. Thus, for example, the statements

```matlab
>> x = 0.0:0.1:10.0;, sine = sin(x);, cosine = cos(x);
>> hdl1 = figure( 'position', [0 0 500 300], 'name', 'Sine Curve' );
>> hdl2 = figure( 'position', [0 400 500 300], 'name', 'Cosine Curve' );
>> figure( hdl1 ), plot( x, sine )
>> figure( hdl2 ), plot( x, cosine )
```

will create two separate windows, each with a suitable title, and then plot the sine curve in the first window and the cosine curve in the second. Both will be visible on the screen at the same time; the second `plot` command will not erase the first window.

### 4.17.4 Customizing Axes

So far, we have used the command `axis` either to set the range on each axis (both in two and in three dimensions) or to establish a square plotting area in each coordinate plane. The command has numerous other features. We mention the following few:

- `axis equal`, which makes equal coordinate intervals of equal length on the screen.
- `axis square`, which makes each axis have the same length.
• **axis normal** (the default), which restores the default size of the plotting area and restores the default auto-scaling on the axes.

• **axis auto** (the default), which stipulates “nice” limits on the axes.

• **axis tight**, which forces the limits on the axes to conform to the actual limits of the data.

• **axis on** (the default), which turns tic marks and labels on for all axes.

• **axis off**, which turns off the axes altogether, including tic marks.

• **axis ticx**, which turns tic marks and labels on for the x axis and off for all other axes. ticy, ticxz, etc. have corresponding responses, displaying tic marks and labels on the specified axes and suppressing them on the unspecified axes. Note that turning off tic marks also turns off labels.

• **axis labelx**, which displays labels on the x axis but suppresses them on all remaining axes (y or yz). Tic marks remain on axes with suppressed labels.

• **axis nolabel**, which suppresses labels on all axes but leaves the axes and the tic marks.

• **box**, which toggles between a border around the plot region (box on) and axes on the left and bottom but not on the top and the right (box off).

• **grid**, which toggles between displaying (grid on) and not displaying (grid off) the full (dotted) grid.

Note that **axis equal** and **axis square** are *not* the same unless the actual ranges of the data on the axes are the same. To expose the differences among normal (the default), equal, image, and square, we execute the statements

```matlab
>> t = [ 0.0 : 2.0*pi/100.0 : 2.0*pi ]; Create vector of values 0 \leq t \leq 2\pi.
>> y = sin(t); Calculate y coordinates for circle.
>> x1 = cos(t) + 1.25;, x2 = x1 - 2.5; Calculate x coordinates for two circles.
>> subplot(2,2,1)
>> plot(x1,y), hold on, plot(x2,y), axis normal, ...
    text(-1.0,0.0,'normal', 'fontsize', 16)
>> set( gca, 'fontsize', 16 )
>> subplot(2,2,2)
>> plot(x1,y), hold on, plot(x2,y), axis equal, ...
    text(-0.5,1.25,'equal', 'fontsize', 16)
>> set( gca, 'fontsize', 16 )
>> subplot(2,2,3)
>> plot(x1,y), hold on, plot(x2,y), axis square, ...
    text(-1.5,0.0,'square', 'fontsize', 16)
>> set( gca, 'fontsize', 16 )
>> subplot(2,2,4)
>> plot(x1,y), hold on, plot(x2,y), axis tight, ...
    text(-2.0,0.0,'tight', 'fontsize', 16)
>> set( gca, 'fontsize', 16 )
>> hold off
```

The resulting display is shown in Fig. 4.28. Notice that only the specification **axis equal** creates axes that show the true geometry of the figure (no stretching along either direction).

### 4.18 Miscellaneous Occasionally Useful Tidbits

In this section, we enumerate features of OCTAVE that are used so infrequently that they are easily forgotten between uses.
4.18 MISCELLANEOUS OCCASIONALLY USEFUL TIDBITS

Figure 4.28: The effect of different axis configurations.

4.18.1 Specifying Directories

Different operating systems use different forms to specify the path that identifies a particular directory, for example, to set the default directory or to identify the location of a file to be read or written. The character separating directories in a path depends on the operating system. In UNIX, that character is a forward slash; in Windows, it is a backslash, though a single forward slash will be properly interpreted even in Windows.

4.18.2 Deleting Variables No Longer Needed

In the course of a long session, numerous no-longer needed variables may accumulate in OCTAVE’s workspace. Sometimes, we may wish to buy a bit of workspace by removing those variables altogether. The function `clear` is provided for this purpose. For example, the statement

```octave
>> clear a b?c d*e
```

will clear the visible variable `a`, all visible three-character variables beginning with `b` and ending with `c`, and all visible variables beginning with `d`, ending with `e` and having any number of characters in between. That is, `?` is a wild card for a single character and `*` is a wild card for any number of characters. The statement `clear` with no arguments will clear all user-defined variables. Execute the statement `help clear` to see a full description, including the several options that are understood by the command.

4.18.3 OCTAVE’s Current Directory

If OCTAVE is launched with a command typed in the `Shell` or `Command` window, the current (or default) directory—the directory into which, by default, files will be written and from which, by default, files will be read—is likely to be the directory from which OCTAVE is launched. If, on
the other hand, OCTAVE is launched by double-clicking ML on an icon, the current directory will typically be the user’s home directory, though that can be changed by a statement in a suitable start-up file (see Section 4.18.5). In the GUI, the identity of the current directory is conveyed in a text-entry box in the toolbar at the top of the window and can be edited by changing the entry in that box, either by browsing in the drop-down menu available in that box, by explicitly typing the desired directory in the box, or by selecting from the folders displayed in the panel immediately below that box. In all interfaces, the identity of the current directory can be displayed with the UNIX-style command `pwd` and changed with the UNIX-style command `cd`. When this directory is unknown or ill-defined, including a full specification of a path in a file name is necessary.

4.18.4 OCTAVE’s Search Path

Whenever an explicit path is not specified for a file that OCTAVE must read (either data file or m-file), the program looks in a succession of directories. By default, that path contains the current working directory (see Section 4.18.3) followed by several directories distributed with OCTAVE, including directories, subdirectories, and subdirectories in

\[ $OCTAVEHEAD/octave-version \]

The command `path` or `path()` with no arguments will display a list of the directories in the current search path. The command

- **addpath** adds one or more directories to the current search path.
- **rmpath** removes one or more directories from the current search path.
- **genpath** replaces the current search path with a new search path containing one or more specified directories.
- **restoredefaultpath** re-establishes the default search path.

See the online help or execute `help path` for more on how to modify the search path.

4.18.5 Customizing OCTAVE

The configuration of OCTAVE can be customized in several ways. First, the system manager can define two system-wide startup files. These files allow the system manager to customize all versions of OCTAVE (the first file in each pair in the footnote) and modify that customization for specific versions (the second file). They can be changed only by the system manager, but they can be opened and viewed by any user. When OCTAVE is launched by any user, the first of these files is executed and then the second is executed.

Additionally, individual users can change the default font and font size, specify additional startup modifiers, and perhaps modify other attributes of OCTAVE in the file `~/.octaverc` in the

\[ $OCTAVEHEAD/octave-4.4.1/share/octave/site/m/startup/octaverc \]  
\[ $OCTAVEHEAD/octave-4.4.1/share/octave/4.4.1/m/startup/octaverc \]  
\[ $OCTAVEHEAD/octave-5.2.0/mingw??/share/octave/site/m/startup/octaverc \]  
\[ $OCTAVEHEAD/octave-5.2.0/mingw??/share/octave/5.2.0/m/startup/octaverc \]

The symbol `$OCTAVEHEAD`, whose value will vary from site to site, must be replaced by the actual path as described in the *Local Guide*. The name and location of these files varies with the version. In OCTAVE 4.4.1 in Windows 10, the files are

- \[ $OCTAVEHEAD/octave-4.4.1/share/octave/site/m/startup/octaverc \]  
- \[ $OCTAVEHEAD/octave-4.4.1/share/octave/4.4.1/m/startup/octaverc \]

In OCTAVE 5.2.0 in Windows 10, with ?? either 32 or 64, they are

- \[ $OCTAVEHEAD/octave-5.2.0/mingw??/share/octave/site/m/startup/octaverc \]  
- \[ $OCTAVEHEAD/octave-5.2.0/mingw??/share/octave/5.2.0/m/startup/octaverc \]

The value of `$OCTAVEHEAD` varies from site to site. See the *Local Guide*.
user’s home directory and the file .octaverc located in the current working directory. These two files are user created. If they exist, the one in the home directory is executed right after the system manager’s files are executed and, finally, the one in the current directory is executed.\textsuperscript{79} This structure forces the user to accept initially whatever the system manager dictates but then to modify those dictates to customize OCTAVE differently for each current directory in which OCTAVE is used. With this sequence, the system manager can enforce features for all users but individual users can override or embellish the system manager’s dictates to match their specific needs.

4.18.6 Restoring OCTAVE’s Initial State

Sometimes, we would like to restore OCTAVE to its initial state. The simplest way to achieve this objective is to exit from OCTAVE and then restart it. Sometimes (though this route is not as secure as exiting from and reentering OCTAVE), it may be sufficient to invoke the OCTAVE command clear -all, which will clear all local and global user-defined variables and all user-defined functions from the symbol table. If we happen to have customized OCTAVE with a one or both personal startup files, then the statement clear -all would have to be followed by an explicit execution of the statements in those startup files, i.e., with the statements

\begin{verbatim}
>> clear -all
>> ~/.octaverc
>> .octaverc
\end{verbatim}

4.18.7 The Command format

Beyond the formats noted in Section 4.3.1, OCTAVE also accepts the statements

\begin{verbatim}
>> format compact
>> format loose
\end{verbatim}

the first of which causes OCTAVE to use less space in output it directs to the screen and the second of which restores the default. Numerous additional specifications of the format can be found by executing the command help format.

4.18.8 Reading Data from the Keyboard

Occasionally, and perhaps especially in M-files, we may wish to prompt for keyboard entry of one or more values by displaying a message on the screen and then read those values from the keyboard. The OCTAVE command input achieves this objective. Its syntax is illustrated in the statement

\begin{verbatim}
>> x = input( 'Enter the value of x: ' )
\end{verbatim}

Here, the argument of the command defines the prompting message and the value entered—which may be a single value or a set of values—will be assigned to the variable \texttt{x}. If the command is terminated with a semicolon, output of the entered value to the screen will be suppressed. If two or more values are to be entered, then the statement requesting the entry would be

\begin{verbatim}
>> pos = input( 'Enter x and y coordinates: ' )
\end{verbatim}

and, separated by space(s) or a comma, the values will be entered in a single line and enclosed in square brackets. In this case, the value assigned to \texttt{pos} will be a vector.

\textsuperscript{79}If the current directory is, in fact, the user’s home directory, the local startup file is read only once.
4.18.9 Writing Data to the Screen

In OCTAVE, many statements not terminated by a semicolon will display output on the screen automatically. If one wishes more control over the format of the output, simply omitting the terminating semicolons will not be enough. The code segments

```octave
disp('This is a message.')  
disp(' t (s) x (cm)')  
disp(a)  
```

illustrate use of the two commands—`disp` and `fprintf`—to send more carefully formatted information to the screen. All will display the desired information without the identifier `ans =`. Note that, in the use of `fprintf`, the file identifier 1 points to the screen and the associated file is always open; it does not require explicit opening or closing. The command `disp` is particularly useful in command files when on-screen messages about progress are desired.

```
disp('This is a message.')  
disp(' t (s) x (cm)')  
disp(a)  
```

Display a text message.

Display the 3 × 2 array a with labels on the columns but without formatting of the values. It will take a bit of experimentation to insert appropriate spaces to line everything up vertically in the output.

```
a11 = num2str(a(1,1), '%4.1f');  
a12 = num2str(a(1,2), '%5.1f');  
str = ['When t = ', a11, ' s, x = ', a12, ' cm.'];  
disp(str)  
```

Format numbers as strings with `num2str`.

Construct output as single string.

Display concatenated string.

```
fprintf( 1, 'When t = %4.1f s, x = %5.1f cm.
')  
```

illustrate use of the two commands—`disp` and `fprintf`—to send more carefully formatted information to the screen. All will display the desired information **without** the identifier `ans =`. Note that, in the use of `fprintf`, the file identifier 1 points to the screen and the associated file is always open; it does not require explicit opening or closing. The command `disp` is particularly useful in command files when on-screen messages about progress are desired.

4.18.10 Exiting from Procedures

Most often, invoked procedures and loops will run to completion and exit normally to the calling program. Sometimes, perhaps in response to a trapped error, it may be desired to exit from a procedure or a loop prematurely. Two statements—`break` and `return`—are available to achieve these ends. The statement `break` is used to exit from a (single) loop to execute the statement following the loop. Suppose, for example, one has a vector `A` containing `n` integers and one wishes to find the square of the first *even* integer in the vector. The coding

```octave
for i = 1:n  
    if 2*fix( A(i)/2 ) == A(i)  
        break;  
    endif  
endfor
result = A(i)^2
```

will step systematically through the components of `A`, testing each to see if it is even, and stopping when it encounters the first even component. At that point, the program exits the for loop and executes the statement following the `endfor`. Time is *not* spent examining entries in `A` that follow that first even entry.

OCTAVE provides the statement `return` to exit prematurely from a running *procedure* or *function*. For example, the coding

```octave
if den == 0  
    disp('ERROR: Division by zero encountered. Execution terminated.' )  
    return  
endif
```

exit from a loop or function. It is a good practice to use `return` to return to the calling program when an error occurs that requires immediate exit (as in the case of division by zero), as it allows the calling program to be interrupted.

OCTAVE provides the statement `return` to exit prematurely from a running *procedure* or *function*. For example, the coding

```octave
if den == 0  
    disp('ERROR: Division by zero encountered. Execution terminated.' )  
    return  
endif
```

OCTAVE provides the statement `return` to exit prematurely from a running *procedure* or *function*. For example, the coding

```octave
if den == 0  
    disp('ERROR: Division by zero encountered. Execution terminated.' )  
    return  
endif
```

```octave
for i = 1:n  
    if 2*fix( A(i)/2 ) == A(i)  
        break;  
    endif  
endfor
result = A(i)^2
```

will step systematically through the components of `A`, testing each to see if it is even, and stopping when it encounters the first even component. At that point, the program exits the for loop and executes the statement following the `endfor`. Time is *not* spent examining entries in `A` that follow that first even entry.

OCTAVE provides the statement `return` to exit prematurely from a running *procedure* or *function*. For example, the coding

```octave
if den == 0  
    disp('ERROR: Division by zero encountered. Execution terminated.' )  
    return  
endif
```

OCTAVE provides the statement `return` to exit prematurely from a running *procedure* or *function*. For example, the coding

```octave
if den == 0  
    disp('ERROR: Division by zero encountered. Execution terminated.' )  
    return  
endif
```

OCTAVE provides the statement `return` to exit prematurely from a running *procedure* or *function*. For example, the coding

```octave
if den == 0  
    disp('ERROR: Division by zero encountered. Execution terminated.' )  
    return  
endif
```
encountered during execution will display the error message and return control immediately to whatever level called the procedure of function.

4.19 References

Those seeking additional information about OCTAVE should consult the help messages that can be brought up

- By typing `help` or `help command` at the OCTAVE prompt,
- By viewing the help manual that will be inserted in the `Documentation` tab on the GUI by selecting 'Help→Documentation→Online' from the menu bar in the GUI,
- By viewing the document that will be displayed in your browser by selecting 'Help→Documentation→Online' from the menu bar in the GUI, or
- By examining the PDF version of the OCTAVE manual accessible at the URL `http://www.gnu.org/software/octave/octave.pdf`. A printed copy of this manual may also be available at your computational facility.

These documents are what we refer to collectively as the OCTAVE manuals. OCTAVE’s website also has an extensive searchable archive of current problems and fixes as well as a decent FAQ. In addition, as is the case with many free programs, there is extensive, but often scattered information available throughout the web and findable by googling appropriate phrases.

4.20 Exercises

4.20.1 Writing OCTAVE Statements

4.1. Write and test OCTAVE statements to create (a) a five-element column vector, (b) an $8 \times 8$ unit matrix, and (c) a $10 \times 10$ matrix all of whose elements are zero except those on the main diagonal (which are all 2) and those on the diagonals just above and just below the main diagonal (which are all $-1$). Search for a route more efficient than laboriously setting each of the 100 elements in the $10 \times 10$ matrix individually.

4.2. Look up the function `sort` both in the on-line help and in the printed OCTAVE manuals. Then, create a vector of your choice and test the use of `sort`, following the pattern illustrated in the documentation. Finally, write in your own words a brief description of what `sort` does.

4.3. Look up the function `lu` both in the on-line help and in the printed OCTAVE manuals, then use the process described therein to solve the linear equations solved by other means in Section 4.3.4.

4.4. Describe and test a sequence of OCTAVE statements that uses a `for/end` loop to evaluate the dot product of two $n$-component vectors $\mathbf{a}$ and $\mathbf{b}$. In essence, you will have to initialize a variable to zero and then, in the loop, successively add to that variable each of the products $a(i) \cdot b(i)$ in turn.

4.5. (a) Describe and test a sequence of OCTAVE statements that uses a `for/end` loop to evaluate $\sum_i a_i$ when the values of $a_i$ are supplied as the elements of the vector $\mathbf{a}$. In essence you will have to initialize a variable to zero and then, in the loop, successively add to that variable each of the elements $a_i$ in turn. (b) Describe and test a sequence of OCTAVE statements that uses the built-in function `sum` to achieve the same end. (c) Explore and then describe the behavior of the built-in function `cumsum`. (d) Explore and describe the behavior of `sum` and `cumsum` when applied to a matrix rather than a vector.
4.6. Write and test an OCTAVE function that accepts two three-component vectors as input and returns a three-component vector containing the cross product of the two input vectors.

4.7. The Fibonacci numbers are generated by picking two starting values and then adding values successively, where each added value is generated by adding the previous two values, e.g., the list [1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89] shows the first several Fibonacci numbers when the first two are both 1. Write a program that asks for the first two numbers and the ultimate length of the desired list and then calculates and displays the corresponding Fibonacci sequence. See math.temple.edu/~reich/Fib/fibo.html for a fascinating discussion of the significance of Fibonacci numbers.

4.20.2 Finding Eigenvalues and Eigenvectors

4.8. Using eig for the main calculation, write and test a function that accepts a symmetric matrix as input and returns a single matrix, each column of which contains an eigenvalue as its first element and the associated eigenvector as its remaining elements.

4.9. Find the eigenvalues and eigenvectors for each of the following matrices:

(a) \[
\begin{pmatrix}
0 & 1 & 0 \\
1 & 0 & 1 \\
0 & 1 & 0
\end{pmatrix}
\]

(b) a 10 × 10 matrix with zeroes everywhere except that all elements on the main diagonal have the value 2.0 and all elements on the sub and superdiagonals have the value −1.0. Search for a route more efficient than laboriously setting each of the 100 elements in the 10 × 10 matrix individually.

(c) Identify the five eigenvectors in part (b) belonging to the lowest five eigenvalues and, for each, plot a graph whose vertical coordinate is the component of the eigenvector and whose horizontal coordinate is the component number. Actually, in the underlying physical context, it would be more appropriate to plot graphs of the result of augmenting these eigenvectors by placing an element 0.0 both before the first element and after the last element in the eigenvector. If, for example, the eigenvectors are in the columns of evecs, then the statement

\[
\text{>> plot( [ 0.0; evecs(:,1); 0.0 ] )}
\]

would plot the requested graph for the first column of evecs—though you should try to improve the appearance of the plot by tampering with the scales, adding labels, .... In particular, the statements

\[
\text{>> x = 0.0 : 1.0/11.0 : 1.0}
\]
\[
\text{>> plot( x, [ 0.0; evecs(:,1); 0.0 ] )}
\]

will produce a graph whose horizontal axis is more suitably labeled. In such a display, you should see something close to the lowest several modes of a vibrating string fixed at both ends!

(d) (Optional) Repeat parts (b) and (c) but with a similarly constructed matrix that is 50 × 50.

(e) (Optional) Use function sort to sort the eigenvalues found in parts (b) and (d) into increasing order and then sort the matrix of eigenvectors to match.

4.10. When a (weak) constant external electric field of magnitude \( F \)—We reserve \( E \) for energy in this exercise.—is imposed on a hydrogen atom, the energies of the states with principal quantum number \( n \) shift from the energies given by the Bohr model by amounts determined by the eigenvalues of the matrix whose elements are \( \langle nlm|eFz|nl'm' \rangle \), where \( l, m, l', \) and \( m' \) range over all possible values of those quantum numbers allowed by the particular value of \( n \). If the states by which the rows
and columns are labeled are ordered \( |2, 0, 0\rangle, |2, 1, -1\rangle, |2, 1, 0\rangle, \) and \( |2, 1, 1\rangle, \) then the matrix for the state \( n = 2 \) is

\[
3e a_0 F \begin{pmatrix}
0 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 \\
-1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}
\]

where \( e \) is the magnitude of the charge on the electron and \( a_0 \) is the Bohr radius. Similarly, if the states by which the rows and columns are labeled are ordered \( |3, 2, 2\rangle, |3, 1, 1\rangle, |3, 2, 1\rangle, |3, 0, 0\rangle, |3, 1, 0\rangle, |3, 2, 0\rangle, |3, 1, -1\rangle, |3, 2, -1\rangle, \) and \( |3, 2, -2\rangle, \) then the matrix for the state \( n = 3 \) is

\[
3e a_0 F \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -9/2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -3\sqrt{6} & 0 & -9/\sqrt{3} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -9/\sqrt{3} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -9/2 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -9/2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]

Find the eigenvalues and eigenvectors of these matrices. The eigenvalues give the energy shifts for the Stark effect for \( n = 2 \) and \( n = 3 \) and the eigenvectors give the linear combinations of the base states (i.e., the states in the absence of the external field) out of which the states in the presence of the field emerge as the field is turned on.

4.20.3 Graphing Scalar Functions of a Single Variable

4.11. Create and test a sequence of OCTAVE statements that will produce a graph that is similar to Fig. 4.6 except that the axes are drawn along the lines \( y = 0 \) and \( x = 0 \).

4.12. (a) Create a vector of 201 elements whose values range from \(-10.0\) to \(10.0\) in equal steps, (b) obtain a graph of the function

\[
y(x) = \frac{x}{a + b(x - c)^2}
\]

over the interval \(-10.0 \leq x \leq 10.0\) with \( a = 1.0, b = 2.0, \) and \( c = 1.0, \) and (c) explore the way the function depends on the three parameters \( a, b, c, \) and (d) write a paragraph describing that behavior in words.

4.13. Consider a circular disk of radius \( a \) lying in the \( xy \) plane with its center at the origin. If the disk carries a uniform charge on its surface, the electrostatic potential at the point \((0, 0, z)\) on the axis of the disk is given by

\[
V(z) = E_0 \left[ \sqrt{a^2 + z^2} - |z| \right]
\]

where \( E_0 \) is a constant. Obtain a graph of \( V(z)/(E_0 a) \) versus \( z/a \).

4.14. The voltage drop across an initially uncharged capacitor in a series RC circuit that is connected at time \( t = 0 \) to a battery is given by the expression

\[
V(t) = V_0 \left( 1 - e^{-t/RC} \right)
\]

Obtain a family of graphs showing \( V(t)/V_0 \) versus \( t \) for various values of \( RC \), and write a paragraph describing these graphs.

4.15. In a Fabry-Perot interferometer, a very large number of waves, each out of phase with the previous one by an amount \( \delta \) and reduced in amplitude by a factor \( r, 0 \leq r < 1, \) interfere. The resulting intensity is proportional to the expression

\[
I(\delta) = \frac{1}{1 - 2r \cos \delta + r^2}
\]
CHAPTER 4. INTRODUCTION TO OCTAVE

Obtain graphs of \( I(\delta) \) versus \( \delta \) for various values of \( r \), and write a paragraph describing these graphs.

4.16. The thin lens equation

\[
\frac{1}{s} + \frac{1}{s'} = \frac{1}{f}
\]

relates object distance \( s \), image distance \( s' \), and focal length \( f \). Here, \( s > 0 \) for an object to the left of the lens, \( s' > 0 \) for an image to the right of the lens, and \( f > 0 \) for a converging lens. With \( s \) ranging from large positive to large negative values, obtain graphs of \( s' \) versus \( s \) for various values of \( f \) (both positive and negative), and write a paragraph describing these graphs.

4.17. According to the special theory of relativity, the mass \( m \), the momentum \( p \), and the kinetic energy \( K \) of a particle moving with speed \( v \) are given in terms of the rest mass \( m_0 \) and the speed of light \( c \) by the equations

\[
m = \frac{m_0}{\sqrt{1 - \beta^2}} \quad ; \quad p = \frac{m_0 u}{\sqrt{1 - \beta^2}} \quad ; \quad K = \frac{m_0 c^2}{\sqrt{1 - \beta^2}} - m_0 c^2
\]

where \( \beta = v/c \). Obtain graphs of \( m/m_0 \), \( p/m_0 c \), and \( K/m_0 c^2 \) versus \( \beta \), superimposing on each a graph of the corresponding non-relativistic expression, and write a paragraph describing these graphs.

4.18. In a vacuum, the transmission and reflection coefficients \( T \) and \( R \) of a dielectric film of thickness \( d \) and index of refraction \( n \) are given by the equations

\[
T = \frac{4n^2}{4n^2 + (n^2 - 1)^2 \sin^2(\kappa d)}
\]

\[
R = \frac{(n^2 - 1)^2 \sin^2(\kappa d)}{4n^2 + (n^2 - 1)^2 \sin^2(\kappa d)}
\]

where \( \kappa = 2\pi n/\lambda \) and \( \lambda \) is the wavelength of the wave in vacuum. Obtain graphs of \( T \) and \( R \) versus \( \lambda/d \) for various values of \( n \) and write a paragraph describing these graphs. Warning: Don’t try plotting too close to \( \lambda = 0 \) since the function \( \sin(\kappa d) \) gives trouble at that point.

4.19. Consider two circular disks, each of radius \( R \), located with their centers on the \( z \) axis such that their planes are parallel to the \( xy \) plane. Let the first disk have its center at the point \((0,0,b/2)\) and the second at the point \((0,0,-b/2)\) so that the disks are separated by a distance \( b \) and the origin is halfway between them. If the top disk carries a uniform, constant charge density \( \sigma \) and the bottom disk carries a uniform, constant charge density \(-\sigma\), the electrostatic potential at the point \((0,0,z)\) is given by

\[
V(z) = \frac{\sigma}{2\epsilon_0} \left[ \sqrt{R^2 + \left(z - \frac{b}{2}\right)^2} - \sqrt{R^2 + \left(z + \frac{b}{2}\right)^2} + \frac{b}{2} \right]
\]

Obtain graphs of \( V(z)/(\sigma R/2\epsilon_0) \) versus \( z/R \) for various values of \( b/R \) and write a paragraph describing these graphs.

4.20. Using data from an experiment you have performed, create a suitable ASCII text file, read the data into OCTAVE, and produce a graph showing the data, error bars, and a theoretical curve. Label the graph completely.

4.21. Journal articles often contain graphs showing data taken by several different scientists regarding the same relationship, with each experimentalist’s contribution marked with a different symbol. Describe an OCTAVE procedure to produce such a graph and, if you have suitable data available (or if you can invent some), demonstrate that your procedure works.

4.22. When a photon of initial energy \( E_0 \) undergoes Compton scattering with an atom of mass \( m \) and is scattered by an angle \( \theta \), the energy of the photon is reduced to

\[
E(\theta) = \frac{E_0}{1 + \xi(1 + \cos \theta)}
\]
where \( \xi = E_0/mc^2 \). Obtain both Cartesian and polar graphs of \( E(\theta)/E_0 \) versus \( \theta \), \( -\pi \leq \theta \leq \pi \), for several values of \( \xi \), and write a paragraph describing these graphs.

4.23. A charged particle moves along the \( z \) axis with speed \( v \). When the particle passes through the origin, the magnitude of the electric field produced by the particle is given by the expression

\[
E(\theta) = \frac{q}{4\pi \epsilon_0 r^2} \frac{1 - \beta^2}{(1 + \beta^2 \sin^2 \theta)^{3/2}}
\]

where \( \theta \) is the polar angle of the observation point, \( r \) is the radial coordinate at that point, and \( \beta = v/c \) (\( c \) is the speed of light). Obtain graphs of \( E/(q/\epsilon_0 r^2) \) versus \( \theta \) for various values of \( \beta \) on the interval \( -\pi \leq \theta \leq \pi \), and write a paragraph describing these graphs.

4.24. The intensity of the interference pattern produced by four slits illuminated by light of wavelength \( \lambda \) when each slit is separated from the next by a distance \( a \) is given by

\[
I(\delta) = \cos^2 \delta (1 + \cos \delta)
\]

where \( \delta = (2\pi a \sin \theta)/\lambda \). Obtain both Cartesian and polar graphs of \( I \) versus \( \theta \) on the interval \( -\pi/2 \leq \theta \leq \pi/2 \) for various values of \( \lambda/a \), and write a paragraph describing these graphs.

4.25. The Planck radiation law gives the expression

\[
u(\lambda, T) = \frac{8\pi ch}{\lambda^5} \frac{1}{e^{h\lambda/kT} - 1}
\]

for the distribution of energy in the radiation emitted by a black body. Here, \( c \) is the speed of light, \( h \) is Planck’s constant, \( k \) is Boltzmann’s constant, \( \lambda \) is the wavelength of the radiation, and \( T \) is the absolute temperature. Using appropriate dimensionless units, plot this function (a) as a function of \( \lambda \) for several \( T \) and (b) as a surface over the \( \lambda T \)-plane. Write a paragraph about the way the peak changes in position, height, and width as \( T \) changes. Hint: Choose a reference wavelength \( \lambda_0 \) arbitrarily and recast the expression in terms of the dimensionless variable \( \Lambda = \lambda/\lambda_0 \). Then, note that \( T_0 = ch/(\lambda_0 k) \) has the dimensions of temperature and re-express the temperature \( T \) in terms of the dimensionless quantity \( \tau = T/T_0 \). (You might find it informative to evaluate \( T_0 \) for \( \lambda_0 = 550 \) nm.) With these changes, the expression to be plotted can be recast in the form

\[
u(\lambda, T) = \frac{1}{\Lambda^5} \frac{8\pi ch}{\lambda_0^5} \frac{1}{e^{h\lambda_0/kT} - 1}
\]

and the question now becomes one of plotting this quantity using the dimensionless variables \( \Lambda \) and \( \tau \).

4.26. A solenoid of length \( L \) and circular cross-section of radius \( a \) lies with its axis along the \( z \) axis and its center at the origin. When the solenoid carries a current, the magnetic field at the point \((0, 0, z)\) on the axis of the solenoid is given by

\[
B(z) = \frac{1}{2} B_0 \left[ \frac{z + L/2}{\sqrt{a^2 + (z + L/2)^2}} - \frac{z - L/2}{\sqrt{a^2 + (z - L/2)^2}} \right]
\]

where \( B_0 \) is the magnetic field at the center when \( a \ll L \), i.e., when the solenoid is effectively infinite in length. Plot graphs showing \( B(z)/B_0 \) (a) as a function of \( z/L \) for various values of \( a/L \), (b) as a surface over the \( (z/L)(a/L) \) plane, and (c) as a contour over the \( (z/L)(a/L) \) plane. Write a paragraph describing these graphs.

4.27. Consider two circular current loops, each of radius \( a \) and lying with its center on and its plane perpendicular to the \( z \) axis. The first loop is centered at the point \((0, 0, b)\) and the second loop is
centered at the point \((0, 0, -b)\). The axial component of the magnetic field at the point \((0, 0, z)\) is given by the equation
\[
B(z) = \frac{1}{2} B_0 \left( a^2 + b^2 \right)^{3/2} \left[ \frac{1}{(a^2 + (z + b)^2)^{3/2}} + \frac{1}{(a^2 + (z - b)^2)^{3/2}} \right]
\]
where \(B_0\) is the magnetic field at the origin. Plot graphs showing \(B(z)/B_0\) as a function of \(z/a\) for various values of \(b/a\), \(b\) as a surface over the \((z/a)(b/a)\) plane, and \(c\) as a contour over the \((z/a)(b/a)\) plane. Write a paragraph describing these graphs.

4.28. In an LRC circuit of resonant frequency \(\omega_0\), the current \(I\) is given as a function of frequency \(\omega\) by
\[
I = \frac{I_0}{\sqrt{1 + Q^2 \left( \frac{\Omega}{\omega} - 1 \right)^2}}
\]
where \(\Omega = \omega/\omega_0\). Plot \(I/I_0\) (a) as a function of \(\Omega\) for various values of the quality factor \(Q\) and (b) as a surface over the \(\Omega Q\)-plane. Write a paragraph describing these graphs.

4.29. A spherical potato of radius \(a\) is taken from the refrigerator at 0 °C and placed in an oven at 200 °C. The temperature \(u(r, t)\) at a point a distance \(r\) from the center of the potato at time \(t\) is given by
\[
\frac{u(r, t)}{u_0} = 1 - 2 \sum_{n=1}^{\infty} \frac{j_0(\beta_n r/a)}{\beta_n J_1(\beta_n)} e^{-\kappa \beta_n^2 t/(ca^2)}
\]
where \(\kappa\) is the thermal conductivity of the potato, \(c\) is its heat capacity per unit volume, \(j_0(x)\) and \(j_1(x)\) are the zeroth- and first-order spherical Bessel functions, and \(\beta_n\) is the \(n\)-th root of \(j_0(x)\), i.e., \(j_0(\beta_n) = 0\). Obtain graphs of \(u(r, t)/u_0\) as a function of \(r/a\) for various values of \(t\). Obtain also a graph of the temperature \(u(0, t)\) at the center of the potato as a function of \(t\) and determine how long it takes the potato to bake if, by being baked, one means that the temperature at the center has risen to 175 °C, i.e., to a value such that \(u(0, t)/u_0 = 0.875\). Hints: (1) Note that
\[
j_0(x) = \frac{\sin x}{x} \quad \text{;} \quad j_1(x) = \frac{\sin x}{x^2} - \frac{\cos x}{x}
\]
Thus, the \(n\)-th root of \(j_0(x)\) is \(\beta_n = n\pi\). (2) Express times in units of \(ca^2/\kappa\) but then, taking the radius of the potato to be \(a = 0.05\) m and taking \(\kappa\) and \(c\) for the potato to be those of water \([\kappa = 0.63\ \text{J/(m s K)}]\), and \(c = 4.2 \times 10^6\ \text{J/(K m^3)}\), determine the unit in which your answers are expressed, both in seconds and in hours. (3) Experiment a bit, but note that the exponential factor decays more rapidly as \(n\) increases, so truncation of the infinite series at some point is probably in order.

4.30. At a particular time, a planet of mass \(M\) is located at the origin in the \(xy\) plane and a moon of mass \(M/3\) is located at a point a distance \(R\) from the planet on the \(x\) axis. The gravitational potential energy of a spaceship of mass \(m\) at the point \((x, y, z)\) is then given by
\[
V(x, y, z) = -\frac{GmM}{\sqrt{x^2 + y^2 + z^2}} - \frac{GmM/3}{\sqrt{(x - R)^2 + y^2 + z^2}}
\]
Using OCTAVE, obtain surface plots and contour maps of this potential energy in the \(xy\) plane (i.e., the plane \(z = 0\)) and in the planes \(z = 0.1R\) and \(z = 0.5R\). Suggestion: Recast the function in dimensionless form by measuring \(x\), \(y\), and \(z\) in units of \(R\) and \(V(x, y, z)\) in units of \(GmM/R\).

4.20.5 Graphing Scalar Functions of Three Variables

4.31. Using either mesh or contour slices, explore at least one of the three-dimensional scalar fields
\[
p_{3,1,0}(x, y, z) = \frac{8}{(27)^{3/2}} \rho^2 \left(1 - \frac{\rho}{6}\right)^2 e^{-2\rho^3/3} \cos^2 \theta
\]
4.20. EXERCISES

4.20.6 Graphing Two-Dimensional Vector Fields

The (gauge) pressure \( p(x, y, z, t) \) inside a cubical box located in the region \( 0 \leq x, y, z \leq a \) is given by

\[
p(x, y, z, t) = A \sin \frac{\pi x}{a} \sin \frac{m\pi y}{a} \sin \frac{n\pi z}{a} \cos \omega_l m n t
\]

where \( l, m, \) and \( n \) are positive integers. Obtain several presentations of the pressure distribution inside this box at \( t = 0 \) and at \( t = \pi/\omega_{l m n} \) for several different values of \( l, m, \) and \( n \).

\[
p_{3,1,1}(x, y, z) = \frac{A}{(2\pi)^3} \rho^2 \left(1 - \frac{\rho}{6}\right)^2 e^{-2\rho/3} \left(1 - \cos^2 \theta\right)
\]
\[
p_{3,2,1}(x, y, z) = \frac{3}{(2\pi)^3} \rho^4 e^{-2\rho/3} \cos^2 \theta \left(1 - \cos^2 \theta\right)
\]
\[
p_{3,2,2}(x, y, z) = \frac{3}{4(2\pi)^3} \rho^4 e^{-2\rho/3} \left(1 - \cos^2 \theta\right)^2
\]
giving the probability density for the hydrogen states \((n, l, m) = (3, 1, 0), (n, l, m) = (3, 1, 1), (n, l, m) = (3, 2, 1), \) and \((n, l, m) = (3, 2, 2)\). These fields are expressed in dimensionless form, where \( \rho \) is the radial coordinate in units of the Bohr radius. In terms of the Cartesian coordinates \( x, y, z \), \( \rho = \sqrt{x^2 + y^2 + z^2} \) and \( \cos \theta = z/\rho \). Hint: To avoid divisions by zero, recast the expressions in terms of \( x, y, z \) explicitly before evaluating any of them numerically.

4.32. The (gauge) pressure \( p(x, y, z, t) \) inside a cubical box located in the region \( 0 \leq x, y, z \leq a \) is given by

\[
p(x, y, z, t) = A \sin \frac{\pi x}{a} \sin \frac{m\pi y}{a} \sin \frac{n\pi z}{a} \cos \omega_l m n t
\]

where \( l, m, \) and \( n \) are positive integers. Obtain several presentations of the pressure distribution inside this box at \( t = 0 \) and at \( t = \pi/\omega_{l m n} \) for several different values of \( l, m, \) and \( n \).

4.20.6 Graphing Two-Dimensional Vector Fields

4.33. Suppose the functions \( F_x(x, y) \) and \( F_y(x, y) \) give the \( x \) and \( y \) components of a vector field at the point \((x, y)\) in the \( xy \) plane. A very crude algorithm for determining the coordinates of points on the field line that starts at the point \((x_0, y_0)\) would entail the operations

\[
x_{old} \leftarrow x_0 \quad \text{Set starting point.}
\]
\[
y_{old} \leftarrow y_0 \quad \text{Choose step.}
\]
\[
ds \leftarrow \text{chosen step size} \quad \text{Choose step.}
\]

\[
\begin{align*}
\text{LOOP} \\
FX & \leftarrow F_x(x_{old}, y_{old}) \quad \text{Start loop to generate steps.} \\
FY & \leftarrow F_y(x_{old}, y_{old}) \quad \text{Calculate components of force.} \\
FM & \leftarrow \sqrt{FX^2 + FY^2} \quad \text{Calculate magnitude of force.} \\
x_{new} & \leftarrow x_{old} + ds*FX/FM \quad \text{Step } dx \text{ along force line.} \\
y_{new} & \leftarrow y_{old} + ds*FY/FM \quad \text{Step } dy \text{ along force line.} \\
\text{Draw line from } (x_{old}, y_{old}) \text{ to } (x_{new}, y_{new}) \\
x_{old} & \leftarrow x_{new} \quad \text{Shift focus to new point.} \\
y_{old} & \leftarrow y_{new} \\
\text{EXIT LOOP WHEN DONE}
\end{align*}
\]

Cast the loop in this algorithm as an OCTAVE procedure \texttt{trace_field} that would be called with a statement like

\[
>> \texttt{trace_field}(x_{old}, y_{old}, ds, n)
\]

where \texttt{xold} and \texttt{yold} convey the desired starting point for a field line, \texttt{ds} conveys the desired step size, and \texttt{n} stipulates the number of steps to be taken before completing the line and returning to the calling program. Though we might ultimately want to provide a means to define the field outside of the tracing procedure, for purposes of this exercise suppose that the field of interest is given by

\[
F_x(x, y) = \frac{x - 1.5}{\sqrt{(x - 1.5)^2 + y^2}} - \frac{x + 1.5}{\sqrt{(x + 1.5)^2 + y^2}}
\]
\[
F_y(x, y) = \frac{y}{\sqrt{(x - 1.5)^2 + y^2}} - \frac{y}{\sqrt{(x + 1.5)^2 + y^2}}
\]

which—in unspecified units—gives the electric field produced in the \( xy \) plane by two point charges, one of relative strength +1 located at \((x, y) = (1.5, 0)\) and the other of relative strength −1 located...
at \((x, y) = -(1.5, 0)\). As a first pass, suppose we are interested in the field only in the region \(-1 \leq x, y \leq 1\) but, once you have your program operating successfully, you can relax that constraint. (Be aware, however, in relaxing the constraint that your algorithm will probably give troubles if you try to plot too close to either of the point charges.) When executed, \texttt{trace\_field} is to draw the field line that starts at the specified point and continues until \(n\) steps have been made. Structure your procedure so that its execution does not erase the current plot, i.e., so that you can execute it repeatedly with different starting points to build up a full map of the field of interest. Optional: Refine the procedure so that

(a) the loop is terminated not only after execution of a fixed number of steps but also when the field line has gone out of bounds, approaches too closely to a singular point, or returns to a point too close to its starting point (closed field line),

(b) it obtains its field definitions from two functions whose names are supplied as arguments when it is called, and/or

(c) it uses a more refined predictor-corrector algorithm in which the values \(x_{\text{new}}\) and \(y_{\text{new}}\) determined above are instead regarded as \(x_{\text{pred}}\) and \(y_{\text{pred}}\), the field is calculated at that predicted point, and then a final step to \(x_{\text{new}}\) and \(y_{\text{new}}\) is made from \(x_{\text{old}}\) and \(y_{\text{old}}\) by using the average of the fields at the points \((x_{\text{old}}, y_{\text{old}})\) and \((x_{\text{pred}}, y_{\text{pred}})\). \textit{Hints:} You may want to use the command \texttt{plot} with the keyword \texttt{nodata} set to \texttt{true} to establish the axes before beginning to trace any field lines.

\textbf{4.34.} Since equipotential curves are perpendicular to the field lines representing the associated (conservative) force field, an algorithm for tracing equipotential curves in two dimensions differs from the algorithm described in the previous exercise only by arranging for the stepping to occur at right angles to the calculated field rather than in the direction of the field. That objective is accomplished by replacing the two statements evaluating \(x_{\text{new}}\) and \(y_{\text{new}}\) in the algorithm in the previous exercise with the statements

\[
\begin{align*}
x_{\text{new}} &\leftarrow x_{\text{old}} - ds*FY/FM \\
y_{\text{new}} &\leftarrow y_{\text{old}} + ds*FX/FM
\end{align*}
\]

Following a pattern similar to that described in the previous exercise, cast the loop in this algorithm as an OCTAVE procedure \texttt{trace\_equipot} that would be called with a statement like

\[
>> \texttt{trace\_equipot( x_{\text{old}}, y_{\text{old}}, ds, n )}
\]

where \(x_{\text{old}}\) and \(y_{\text{old}}\) convey the desired starting point for the equipotential curve, \(ds\) conveys the desired step size, and \(n\) stipulates the number of steps to be taken before completing the curve and returning to the calling program. Then, explore the equipotential curves depicting the field on which the previous exercise focussed.

\textbf{4.35.} Sometimes, interest can be focussed on the \textit{magnitude} of the vector field rather than on the full field. A command like

\[
\texttt{MAG} = \texttt{sqrt( Hx.^2 + Hy.^2 )}
\]

for example, will generate an array \(\texttt{MAG}\) containing the magnitudes of the magnetic field at each available grid point. This array then conveys a \textit{scalar} field that can be displayed by any of the methods described in Section 4.13. Explore the magnetic field given in Eq. (4.16) in this way.

\textbf{4.36.} Read the OCTAVE manuals about the procedure \texttt{quiver3}. Then use \texttt{quiver3} to explore the velocity field

\[
v(x, y, z) = \omega(-y \mathbf{i} + x \mathbf{j}) + \alpha \mathbf{k}
\]

(which happens not to depend on \(z\)) for various values of \(\omega\) and \(\alpha\). Then write a paragraph or two describing the field. \textit{Reassurance:} Diagrams of vector fields in three dimensions are not easy to fathom. Do not be dismayed if the pictures you generate are initially mysterious. \textit{Hint:} In
establishing a scale, you may find it useful to assume a length scale \( a \), recast the field in the form

\[
v = a\omega \left( -\frac{y}{a} \hat{i} + \frac{x}{a} \hat{j} \right) + \alpha \hat{k} \quad \Rightarrow \quad \frac{v}{a\omega} = \left( -\frac{y}{a} \hat{i} + \frac{x}{a} \hat{j} \right) + \frac{\alpha}{a\omega} \hat{k}
\]

and try to draw graphs of the field \( \frac{v}{a\omega} \) as functions of the parameter \( \frac{\alpha}{a\omega} \) in the space whose axes are labeled \( x/a, y/a, \) and \( z/a \).

### 4.20.7 Animation

#### 4.37.

The transverse motion of a flexible string of length \( l \) lying nominally along the \( x \) axis and fixed at both ends can be expressed as the superposition

\[
y(x, t) = \sum_{n=1}^{\infty} A_n \sin \left( \frac{n\pi x}{l} \right) \cos \left( \frac{2\pi nt}{T} \right)
\]

of its normal modes of oscillation. Here, \( A_n \) is the amplitude of the \( n \)-th harmonic (and may be negative to convey a 180° phase shift relative to a mode with positive amplitude) and \( T \) is the period of the fundamental mode of oscillation. In particular, the shape of the string at time \( t = 0 \) is given by

\[
y(x, 0) = \sum_{n=1}^{\infty} A_n \sin \left( \frac{n\pi x}{l} \right)
\]

Measuring \( x \) in units of \( l \) and \( t \) in units of \( T \), generate animated displays by writing a procedure that will accept as input a vector giving the amplitudes of the first fifteen harmonics and produce a continuously running display showing the motion of the string when its initial shape is defined by those amplitudes. Test your program with a variety of sets of amplitudes, including but not limited to the first several harmonics by themselves. For example, when the string is pulled aside at its center and released from rest, the amplitude of the first several harmonics will be

\[
1.0, 0.0, -0.111111, 0.0, 0.04, 0.0, -0.0204082, 0.0, 0.0123457, 0.0, -0.00826446, 0.0, 0.00591716, 0.0, -0.00444444
\]

and, when it is pulled aside very near to one end, the amplitude of the \( n \)-th harmonic will be \( 1/n \).

#### 4.38.

The displacement of a string supporting a transverse wave is given by \( u = f(x, t) \) where \( f(x, t) \) is a given function of \( x \) and \( t \). Develop a general procedure to animate this wave propagation (i.e., for showing a sequence of images created by graphing \( f(x, t) \) as a function of \( x \) for a succession of values of \( t \)) and write an appropriate M-file to accomplish this task. Try your procedure with the two functions

\[
f(x, t) = e^{-(x-ct)^2} \quad \text{and} \quad f(x, t) = \sin(x)\cos(t)
\]

but feel free to invent others of your own choosing. Here, we suppose that \( x \) and \( t \) have been cast in appropriate dimensionless units.

#### 4.39.

The displacement of a square membrane extending over the region \( 0 \leq x, y \leq a \) when it is oscillating in its \( m, n \) normal mode is given by

\[
u(x, y, t) = A \sin \left( \frac{m\pi x}{a} \right) \sin \left( \frac{n\pi y}{a} \right) \cos (\omega_{mn} t)
\]

where \( \omega_{mn} = (c\pi/a) \sqrt{m^2 + n^2} \), \( c \) being the speed of propagation of the waves in the membrane. Write a function M-file to animate the motion of this membrane for a user-selected mode (user-selected values of \( m \) and \( n \)). Suggestions: (1) Express \( x \) and \( y \) in units of \( a \) and \( u \) in units of \( A \) so the graphs you produce will show \( u/A \) above the \((x/a)(y/a)\) plane. (2) Review the discussion in the last paragraph of Section ??.
4.20.8 Supplementary Exercises

4.40. Write a program to (a) accept as input an arbitrary two- to four-digit integer with at least two different digits, (b) rearrange the four digits to create (i) the largest and (ii) the smallest integers from the four digits, (c) subtract the smallest number from the largest numbers, and (d) repeat steps (b)–(c) until the same number occurs twice in succession. As a guard against an infinite loop, limit the number of iterations to 10. You should find, probably with some surprise, that the end result is always 6174, known as a Kaprekar Constant, and that no more than seven iterations will be required. Note: If the input number or any result after you complete step (c) has only three digits, add a leading zero before moving to the next iteration. Note: If you code appropriately, the adding of zeroes may happen automatically. Optional: Explore whether numbers with three or five digits exhibit a similar property and, if they do, identify the corresponding Kaprekar Constant. Google “Kaprekar Constant” to locate several articles on this and similar interesting numbers.
Chapter 8

Introduction to MATHEMATICA

Note: All program (.m) files referred to in this chapter are available in the directory $HEAD/mathematica, where (as defined in the Local Guide) $HEAD must be replaced by the appropriate path for your site. At some sites, this directory or some other directory containing these files may also have been placed in Mathematica’s default search path. If so, the files can be found by Mathematica without explicit specification of a path. Otherwise, you will have to use the full path to copy them into your default directory to access them.

Conceived and developed by Stephen Wolfram and first released to the public in June of 1988, Mathematica® is a large program for manipulating expressions symbolically.1 Mathematica can perform symbolic algebra; evaluate derivatives, integrals, and Taylor series; solve algebraic and differential equations; manipulate complex numbers; find eigenvalues and eigenvectors; produce two- and three-dimensional graphs; and accomplish numerous other tasks. In this chapter, we describe some ways by which expressions can be defined for Mathematica and then illustrate some of the statements by which Mathematica can be instructed to manipulate these expressions. Users should also work through the on-line help messages within Mathematica itself and, in particular, look at the examples with which most on-line help messages conclude (Section 8.2). Further details can be found in printed and on-line Mathematica documentation (Section 8.19). We shall refer to this documentation collectively as the Mathematica manuals.

Two user interfaces to Mathematica are provided, a text-based interface (TBI), which gives direct access to the Mathematica kernel, and a notebook interface (NBI), which displays equations in a more elegant form, offers assistance for those who have trouble remembering the commands, and has a notebook capability to facilitate documentation. In essence, the TBI gives the user direct access to the Mathematica kernel while the NBI is simply a front-end that facilitates construction of statements that the NBI then formulates properly and passes on to the kernel, which may actually be running on a different (and perhaps faster) computer than is used for the front-end. In this book, we assume that the user is working in the NBI. We will, however, describe the structure of input statements without addressing the ways in which the NBI facilitates assembling them,2 so the statements we describe can equally well be submitted to the TBI. We shall present Mathematica output with the approximate appearance it will have in the NBI.

Except where occasionally noted otherwise, for the rest of this book, information provided describes the NBI explicitly and may not accurately reflect the behavior of the TBI.

1 Mathematica is a registered trademark of and also a commercially available program produced and marketed by Wolfram Research, Inc., of Champaign, Illinois. (See Appendix Z for full contact information.) Its use at any particular site is subject to the provisions of whatever license that site has negotiated with Wolfram Research, Inc. The terms of that license are explained in the Local Guide.

2 See the Mathematica manuals for information about how to use templates, palettes, and other aids available in the NBI.
8.1 Beginning a *Mathematica* Session

Detailed instructions for initiating a session with *Mathematica* will be found in the *Local Guide*. In general, *Mathematica* will be started either (1) by typing the command `mathematica` for the NBI (or `math` for the TBI) at the prompt from the operating system,\(^3\) (2) by double-clicking ML on an appropriate icon on the desktop, or (3) by selecting 'Mathematica' (or an equivalent phrase) from a menu.\(^4\) Presently, an introductory screen offering a number of options will appear. Opening the New Document drop-down menu and selecting the option 'Notebook' produces a blank notebook titled 'Untitled-1' and a cursor—but no prompting character(s). We must from the outset be aware that

1. Labels, e.g., `In[12]:=` for input and `Out[12]=` for output, will appear after a statement has been submitted to *Mathematica*.

2. Internally, *Mathematica* is case sensitive, i.e., *Mathematica* treats upper and lower case letters as distinct. By convention, built-in *Mathematica* commands use names whose initial letter is capitalized.

3. By default, *Mathematica* formats input statements using *Wolfram Language input*. Other formats can be selected by opening the menu accessed by the '+' sign that appears at the top of the notebook.

4. Lines of *Mathematica* code can be terminated in any of three ways, depending on whether we want to

   (a) submit the statement for execution and instruct *Mathematica* to display its response, in which case we terminate the statement by typing (SHIFT/ENTER) on the main keyboard or by typing (ENTER) on the numeric keypad.\(^5\)

   (b) submit the statement for execution and instruct *Mathematica* to suppress explicit display of its response, in which case we terminate the statement with a semicolon followed by (SHIFT/ENTER) on the main keyboard or (ENTER) on the numeric keypad.

   (c) continue entering the statement on a new line, in which case we terminate the line by typing simply (ENTER), but we must be clever in exploiting this feature. *Mathematica* will expect additional input only if the part entered before typing (ENTER) is incomplete. To enter the statement `a + b + c` on two lines, for example, we could not break it after the `b` but we could break it after either of the `+` signs.

5. Two or more statements can be placed in a single physical line by separating the statements with semicolons, though output will be suppressed for all but the last statement in the line (unless it is followed by a semicolon).

6. To abort execution of a statement by the *Mathematica* kernel and return *Mathematica* to a state from which new statements can be entered, either type (ALT./.) or, alternatively,

   - Select the block containing the statement whose execution is to be terminated by clicking ML on the corresponding bracket at the right end of the screen,
   - Open the Evaluation menu from the menu bar across the top of the notebook, and
   - Select ‘Abort Evaluation’ from that menu.

   There may be a short delay during which *Mathematica* will clean up its workspace before actually “prompting” for fresh input.

---

\(^3\)In this command to the operating system, case may be important.

\(^4\)Access to *Mathematica* is controlled by a license manager. When *Mathematica* is started, the license manager checks to make sure that a license is available. If no license is available (or the license manager happens not to be running), the desired session will not start. Restarting a stopped license manager requires action by the system administrator.

\(^5\)On some keyboards the (ENTER) key will be labeled (RETURN).
7. Typing the statement Quit[] whenever Mathematica is ready for input cleans Mathematica’s workspace and reinitializes its internal statement counter but does not close the notebook or return control to the operating system. To exit altogether from Mathematica, select ‘Exit’ from the File menu or click ML on the ‘×’ button in the upper right corner above the Mathematica menu bar. If the notebook contains unsaved content, a pop-up box will give you the option of saving that content before Mathematica actually returns control to the computer’s operating system.

To close the current notebook without exiting from Mathematica, select ‘Close’ from the File menu. If, after closing a notebook, you open a new notebook by selecting ‘New→Notebook’ from the File menu or ‘Notebook’ from the Mathematica opening screen, statement numbers in that new notebook will continue from where they left off in the closed notebook. To reinitialize the internal statement counter, execute the statement Quit[], preferably before closing the first notebook but alternatively right after opening the new notebook.

9. The Mathematica command Run[ "OSCommand" ] effects temporary escape from Mathematica to the underlying operating system (OS) to effect execution of OSCommand by that operating system. Control is then returned to Mathematica. The resulting output (if any) is displayed in the command window from which Mathematica was launched.6

10. Although the feature can be disabled, by default some types of error detected by Mathematica’s front end will produce an audible beep. For further information about that beep when it occurs, select ‘Why the beep?’ from the HELP menu.

---

### 8.2 On-Line Help

Mathematica provides extensive on-line help. Users can enter a hypertext browser by selecting ‘Wolfram Documentation’ (or some equivalent phrasing) from the HELP menu. The browser can be used to search for an item by entering the name of the item in the text entry box at the top of the browser and then clicking ML on the magnifying glass next to the text entry box; presently, a description of the item will appear. Alternatively, we can navigate the hierarchy of help messages by opening various submenus represented by colored boxes in the window that pops up when ‘Wolfram Documentation’ is selected.

Alternatively, information about a particular item can be obtained simply by typing a question mark ?—which must be the first character on the line—followed by the name of the item about which information is sought (e.g., ?Sin) whenever Mathematica is prepared to accept input. This action brings up an inline brief description of the specified item, including a link—the symbol >>—whose selection will bring up the full documentation for the item sought. Beginning a line with two question marks, e.g. ??Sin, appends to the message produced by one question mark a statement about the attributes assigned to the specified symbol.

Finally, when a specific Mathematica symbol is highlighted in a notebook, selecting ‘Find Selected Function’ from the HELP menu will bring up the full documentation for that function.

Items in the on-line help library that describe specific Mathematica commands typically include not only a description of the function and syntax of the command but also examples of the proper use of the command and links to further details and related topics at other points in the on-line messages. Many of the examples contain active statements that can be edited and reexecuted from within the help message. Your attention is drawn particularly to the on-line help messages for the commands D, Integrate, Simplify, Expand, Factor, Solve, Sum, and DSolve.

---

6In the NBI on any platform of course, this feature is unnecessary, since we can easily use resources of the desktop to execute other programs without exiting from Mathematica. Indeed, any output produced by Run[ "OSCommand" ] executed from the NBI will be displayed in a pop-up Shell window which, however, will promptly disappear when the command ends.
8.3 Basic Entities in Mathematica

At the very beginning, we point out that Mathematica distinguishes among a variety of different entities, including

- **expressions**, **lists**, and **character strings**, which are the primary quantities with which Mathematica works;
- **commands** and **functions**, which accept zero, one, or more arguments as input and return various outputs, thereby providing the means by which the objects enumerated in the previous item are manipulated;
- **system variables**, e.g., `$HomeDirectory`, `$IgnoreEOF`, `$Packages`, `$Path`, `$RecursionLimit`, `$TextStyle`, and `$TopDirectory`, whose values—default or user-specified—control the behavior of various system functions or record various pieces of information about the session. Some of these variables are maintained by Mathematica and are beyond the reach of individual users; others can be customized to the user’s specific ends.
- **constants**, including Pi for \(\pi\), I for \(\sqrt{-1}\) (which, except in exponents, will be printed in Mathematica output with a somewhat thicker vertical line but which we will display in output more simply as the lower case Roman character \(i\)), \(\text{Infinity}\) for \(+\infty\), \(E\) for the base of the natural logarithm (which, except in exponents, will be printed in Mathematica output with an extra vertical line but which we will display in output more simply with the lower case Roman character \(e\)), and `True` and `False` for the two logical values. **special symbols**, including \(+\) (addition), \(-\) (subtraction), \(*\) (multiplication), \(/\) (division), \(:=\) (function definition), \(\rightarrow\) (perform replacement), \(==\) (equality), and \(-\rightarrow\) (transformation rule).
- **options**, e.g., `Assumptions`, `PlotRange`, `TextStyle`, and `TargetFunctions`, which may appear among the arguments to functions, may be assigned values, and—when present—modify the output produced by the functions. A full list of the options available, including the default value of each, for a particular function will be displayed when the statement `Options[ Function ]`, e.g., `Options[ Plot ]` is executed.

Consistent with Mathematica structure, many of these entities are represented internally in the standard form consisting of a head followed by arguments in the form of a list. The command `FullForm`, which will be introduced in Section 8.8.2, provides a convenient means to display the internal form of these—and many other—Mathematica objects.

8.4 Variable Names in Mathematica

Legal variable names begin with an alphabetic character and cannot coincide with reserved words like `True`, `Sin`, and `Factor`\(^7\) but are otherwise unrestricted. They can contain alphabetic and numeric characters but not the underscore character.\(^8\) Within reason, they can be of any length.\(^9\) Lower-case and upper-case letters are treated as distinct. Typing \(\text{ESCAPE}\) \(\text{gamma}\) \(\text{ESCAPE}\) (or \(\text{\\backslash gamma}\)) in an input line will result in the display of the lower-case or upper-case Greek letter both in the input line and in the associated output line, e.g., \(\text{ESCAPE}\)\(\text{gamma}\)\(\text{ESCAPE}\) (or \(\text{\\backslash gamma}\)) in an input line will produce \(\gamma\) in the input and output lines while \(\text{ESCAPE}\)\(\text{Gamma}\)\(\text{ESCAPE}\) (or \(\text{\\backslash CapitalGamma}\)) will produce \(\Gamma\).

Note particularly that, in almost all situations, Mathematica will assume by default that any variable to which no explicit value has been assigned represents a complex quantity.

\(^7\)Since the names for built-in Mathematica objects (commands, special variables, options, ...) begin with an upper-case letter, avoiding upper-case letters in user-defined variable names will avoid naming conflicts.

\(^8\)This character is reserved in Mathematica for another purpose, to be outlined in Section 8.6.

\(^9\)The actual maximum length depends on the operating system, but is quite large even in the most restrictive case.
8.5 Expressions in *Mathematica*

Expressions are written in *Mathematica* using a syntax very much like that used in ordinary algebra and are supplied merely by typing the expression at *Mathematica*’s prompt for input,\(^{10,11}\) e.g.,

\[
\text{In}[12]:= \text{x + a} \times \text{Cos}[\text{b} \times \text{x}]
\]

to which, when the expression is entered by typing ⟨SHIFT/ENTER⟩,\(^12\) *Mathematica* will respond\(^{13}\)

\[
\text{Out}[12]= \text{x + a} \times \text{Cos}[\text{b} \times \text{x}]
\]

The symbols +, −, *, /, and ^ are used for addition, subtraction, multiplication, division, and exponentiation.\(^{14}\) Except for those in quoted strings,\(^{15}\) spaces are ignored and may be freely inserted for legibility. Functions available for use in expressions include

- General functions, e.g., Abs[], N[], Trunc[], Round[], Sqrt[], Sign[], Random[], and \[].
- Trigonometric, hyperbolic and exponential functions and their inverses, e.g., Sin[], Cos[], Tan[], ArcSin[], ArcCos[], ArcTan[], ArcTan[], Sinh[], Cosh[], Tanh[], ArcSinh[], ArcCosh[], ArcTanh[], Exp[], and Log[]. Note that each function name is capitalized and the argument(s) of the function is (are) enclosed in square brackets.
- Special functions, e.g., Bessel and modified Bessel functions, Airy functions, elliptic integrals, gamma and beta functions, the error function, Legendre polynomials, Hermite polynomials, and Laguerre polynomials.

Full details on these and other functions, including information about the form of their arguments, will be found in the *Mathematica* manuals. A few will be introduced in the remainder of this chapter.

Internally, *Mathematica* regards a simple expression to be composed of a head, which conveys the operator involved, followed by a list of the operands known individually as part one, part two, ..., (The head can also be referred to as part zero.) Thus, for example, the two expressions

\[ a + b + c + d \quad \text{and} \quad abc \]

are stored internally as the expression

\[
\text{Plus}[\text{a}, \text{b}, \text{c}, \text{d}] \quad \text{and} \quad \text{Times}[\text{a}, \text{b}, \text{c}]
\]

\(^{10}\)In this section, we use non-sequential statement numbers to emphasize that many statements may intervene between those illustrated. These statements will all be executable as written, but your statement numbers will almost certainly differ from those presented here.

\(^{11}\)In the TBI, the characters In[12]:= will be displayed as a prompt before the statement is typed. In the NBI, no prompt is displayed; the characters—best called a label in this case—will be added after the statement has been submitted for execution.

\(^{12}\)See item 4 in Section 8.1 for alternatives. We shall make no further mention of the need for an appropriate termination to each statement.

\(^{13}\)Note that we will occasionally rearrange the display of *Mathematica* output lines to improve legibility. No substance is lost in that rearrangement.

\(^{14}\)Mathematica actually allows omission of the symbol * for conveying multiplication. Thus, the statement \(x \times a \times \text{Cos}[b \times x]\) is fully equivalent to the one illustrated in line In[12]. Spaces separating variables to be multiplied are then critical, since \(xy\) would be interpreted as the two-character name of a (single) variable, not as the product of two variables \(x\) and \(y\). While recognizing the simpler appearance of statements in which the asterisk is omitted, we will throughout this book nonetheless note the multiplication explicitly.

\(^{15}\)and those separating entities to be multiplied; see previous footnote.
respectively. More complicated expressions are stored as nested simple expressions. Thus, for example, the expressions

$$a + b - c + d \quad \text{and} \quad a + b(c + d) + e$$

have the internal representations

$$\text{Plus}[\ a, b, \text{Times}[-1, c], d \ ] \quad \text{and} \quad \text{Plus}[\ a, \text{Times}[b, \text{Plus}[c, d]], e \ ]$$

respectively.

As a final example, which anticipates a later need, suppose the expression of interest is the product of three binomial factors, e.g.,

$$(x + 3)^2 (x - a)(x + a)$$

For this expression, we would generate Mathematica’s internal representation in several steps finding

$$\text{Times}[\ (x+3)^2, x-a, x+a \ ]$$

$$\text{Times}[\ \text{Power}[x+3,2], \text{Plus}[-a,x], \text{Plus}[a,x] \ ]$$

$$\text{Times}[\ \text{Power}[\text{Plus}[3,x], 2], \text{Plus}[\text{Times}[-1,a], x], \text{Plus}[a, x]]$$

Here, the first line shows the head and parts 1, 2, and 3 of the original expression while the second line reveals that part 1 of the original expression is itself a structure having a head (Power: the head of part 1 of the original expression), part 1 (x+3; part 1 of part 1), and part 2 (2; part 2 of part 1). Further, the third line reveals that some of the parts of the parts themselves have parts. Thus, for example, part 2 of part 1 of part 1 of the original expression is x while part 1 of part 2 of part 2 of the original expression is -1. These relationships can be involved and confusing, but some manipulations depend critically on being able to identify the parts, subparts, subsubparts, and ... of a complicated expression.\footnote{The third line conveys the representation actually used in Mathematica. Unfortunately, Mathematica does not always store the various parts in the same order in which they were entered. After all, the expressions \text{Plus}[a,b,c] and \text{Plus}[b,a,c] are absolutely equivalent in value. However an expression is entered, Mathematica chooses a standard order to facilitate subsequent manipulations. Presently, we shall encounter the Mathematica command \text{FullForm}, which permits us to instruct Mathematica to display its internal representation.}

8.6 Assigning Values to Variables; Defining Functions

Especially when elaborate expressions are involved, we may want to refer to these expressions after they have been entered without having to retype them. Mathematica provides for this shorthand in two ways. First, as a session proceeds, Mathematica maintains an internal record of the contents of each Out line, identifying each with the number \( n \) that appears in square brackets after the word Out. The item at line \text{Out}[n] can then be referred to in subsequent statements either by the label \text{Out}[n] or by the label \%n. For example, a later request that Mathematica square the expression on line \text{Out}[12] in the previous section could be communicated to Mathematica either with the statement

\begin{verbatim}
In[18]:= %12^2
Out[18]=(x+a Cos[b x])^2
\end{verbatim}
8.6. ASSIGNING VALUES TO VARIABLES; DEFINING FUNCTIONS

In[19]:= Out[12]^2
Out[19]=(x + a Cos[b x])^2

Alternatively, if we anticipate frequent reference to a particular expression, we may wish to use the assignment operator ‘=’ to assign or bind the expression to a variable. The statement,

In[20]:= bigexp = x + a * Cos[b*x]
Out[20]= x + a Cos[b x]

for example, associates the variable bigexp with the expression following the assignment operator =. Subsequently, whenever that variable is included in an expression, it will automatically be replaced by the expression to which it has been bound, e.g., had we bound the expression to the variable bigexp, we might later have written

In[30]:= bigexp^2
Out[30]=(x + a Cos[b x])^2

The wisdom of choosing variable names with mnemonic significance should be obvious.

Once a name has been assigned, we can at any time remind ourselves of the value bound to that variable simply by asserting its name in a statement like

In[47]:= bigexp
Out[47]= x + a Cos[b x]

or, alternatively, by interrogating the system with a statement like

In[48]:= ?bigexp
Global 'bigexp
bigexp = x + a Cos[b x]

Further, if we can’t remember whether we have already assigned a value to a particular variable, we can invoke the statement

In[75]:= Names[ "bigexp" ]
Out[75]= { bigexp }

to find out.\footnote{The statement will return a null list ({} if no value has been assigned to the specified variable. Note that the wildcard * can be used in the string. Further, the statement Names[] will return a (long) list of all the symbols known, including those that Mathematica defines for its own purposes.} Finally, when we have no further need for a particular variable and wish to tell Mathematica to forget that we had once bound a value to it, we could enter either of the statements

In[97]:= Clear[ bigexp ]
In[98]:= Remove[ bigexp ]

the first of which removes any value or definition assigned to the named variable but leaves the variable (and its attributes) known to Mathematica\footnote{Alternatively, the shorter statement bigexp=. will achieve the same end.} while the second of which completely deletes not only any assigned value or definition but also the name itself so the variable is no longer known to Mathematica in any way.
When special functions of significance to a particular session would be useful, we might exploit the Mathematica operator := and the underscore suffix to a variable name to append one or more user-defined functions to the built-in functions. For example, the statements

\begin{verbatim}
In[123]:= spectrig[x_] := 1 + Sin[x]^2
In[124]:= mode[x_,y_] := Sin[Pi*x] * Sin[2*Pi*y]
\end{verbatim}

define the functions \texttt{spectrig}(x) and \texttt{mode}(x,y) for subsequent use just as we would use any of Mathematica's built-in functions. Further, if we need to be reminded of the function definition given to a variable, we need simply enter statements like

\begin{verbatim}
In[170]:= ?spectrig
In[171]:= ?mode
\end{verbatim}

Finally, when we have no further need for a particular function and wish to tell Mathematica to forget that we had once associated a function with a particular variable, we would enter an appropriate one or two of the statements

\begin{verbatim}
In[201]:= Clear[ spectrig ]
In[202]:= Remove[ spectrig ]
In[203]:= Clear[ mode ]
In[204]:= Remove[ mode ]
\end{verbatim}

to remove the function definitions or the definitions and their names from Mathematica's “memory”.

Within Mathematica, some potential variable names are protected. The names \texttt{Pi} (for \(\pi\)), \texttt{EulerGamma} (for Euler's constant, 0.577216), and \texttt{Sin} (for the sine function), for example, are protected. An attempt to assign a value to a protected variable will fail, as illustrated in the statement

\begin{verbatim}
In[227]:= Pi = 4.9
... Set: Symbol \pi\ is Protected.
Out[227]= 4.9
\end{verbatim}

though the content of the message may vary with the version of Mathematica. Further, in the illustrated example, both the symbol ‘...’ and the word ‘Set’ are hot keys that give access to additional information.

In rare circumstances, we may wish to use the (protected) symbol anyway. If so, we must unprotect it before using it in the new context, as illustrated in the statements

\begin{verbatim}
In[240]:= Unprotect[ Pi ]; Pi = 4.9
\end{verbatim}

Once we are finished with the non-standard use, we can restore the original status of the variable with the statements

\begin{verbatim}
In[256] := Pi = .; Protect[ Pi ];
\end{verbatim}

\textsuperscript{19}This suffix, which is used only on the left-hand side of a function definition, flags the variable so that Mathematica will understand that whatever expression, however complicated, appears as the argument of the function is to be substituted for the variable without the suffix on the right-hand side in the function definition.

\textsuperscript{20}Note the use entirely of lower-case letters in naming the user-defined functions. By adhering to this convention, we avoid conflict with Mathematica's reserved words.
where the first statement removes the non-standard definition, restoring the default value, and the second statement reestablishes the protection of the variable. Tampering with specially defined variables and other reserved words, however, is risky, and should be undertaken only occasionally and only for good reason.

### 8.7 Creating and Examining Lists

We have already mentioned in Section 8.3 that Mathematica works with several different types of object, and we have in Sections 8.5 and 8.6 described how to create expressions and assign values to variables. To simplify the quick tour of selected capabilities to be presented in Section 8.8, we include here a description of how lists and lists of lists can be created and examined.

A list is an ordered, comma-separated sequence of elements that is enclosed in braces \{\ldots\}. Lists are created with statements like

\begin{verbatim}
In[301]:= lst1 = {a, b, a, d, e}
Out[301]= {a, b, a, d, e}
In[302]:= lst2 = {"David", 5.00, a+b}
Out[302]= {"David", 5.0, a+b}
\end{verbatim}

An entire list can be referred to by its name (e.g., \texttt{lst2}), while a particular element in the list can be referred to by placing its index in double square brackets after the name of the list (e.g., \texttt{lst2[[2]]}). \footnote{Alternatively, a particular element of a list can be extracted with the function \texttt{Part} to be introduced presently. The element \texttt{lst2[[2]]} can also be identified as \texttt{Part[lst2, 2]}.}

Mathematica contains a few special commands to facilitate the creation of commonly needed lists, including the statements

\begin{verbatim}
In[303]:= Range[5]
Out[303]= {1, 2, 3, 4, 5}
In[304]:= Range[5, 10]
Out[304]= {5, 6, 7, 8, 9, 10}
In[305]:= Range[1, 2, 1/4]
Out[305]= \{1, 5/4, 3/2, 7/4, 2\}
\end{verbatim}

Create a list containing values ranging from 1 to 5 in steps of 1.
Create a list containing values ranging from 5 to 10 in steps of 1.
Create a list containing values ranging from 1 to 2 in steps of \(\frac{1}{4}\).

We draw your attention also to the command \texttt{Table}, which can be used to create a list containing values that can be calculated from a formula. For example, the statement

\begin{verbatim}
In[306]:= Table[n^2, {n, 4}]
Out[306]= \{1, 4, 9, 16\}
\end{verbatim}

creates a list containing the squares of the integers from 1 to 4 while the statement

\begin{verbatim}
In[307]:= Table[Sin[Pi*n/2], {n, 6}]
Out[307]= \{1, 0, -1, 0, 1, 0\}
\end{verbatim}

creates a list of the values of sin\((n\pi/2)\) for \(n\) ranging in integer steps from 1 to 6.

Lists of lists are also easily created. We simply place several lists, each of which is enclosed in braces, as the elements of a larger list. The statement

\begin{verbatim}
In[308]:= \texttt{lst2} = {{"David", 5.0, a+b}, {"John", 6.0, a+b}}
Out[308]= \{\{"David", 5.0, a+b\}, \{"John", 6.0, a+b\}\}
\end{verbatim}
In[308]:= lstlst = { {1,2,3}, {4,5,6}, {7,8,9} }

for example, produces the output

Out[308]= { {1, 2, 3}, {4, 5, 6}, {7, 8, 9} }

In essence, the statement at In[308] creates a list of three elements, each of which is itself a list of three elements. The elements of this list of lists can be extracted with statements like

In[309]:= lstlst[[2]]

Out[309]= {4, 5, 6}

Extract second element of lstlst.

In[310]:= lstlst[[2,3]]

Out[310]= 6

Extract third element of second element of lstlst.

Further, since a list of lists can be viewed as a two-dimensional matrix, Mathematica provides the statement

In[311]:= lstlst // MatrixForm

Out[311]//MatrixForm =

\[
\begin{bmatrix}
1 & 2 & 3 \\
4 & 5 & 6 \\
7 & 8 & 9 \\
\end{bmatrix}
\]

or MatrixForm[lstlst].

to display the list of lists as a matrix. We note that the element lstlst[[1]] is the first row of the matrix, the element lstlst[[2]] is the second row of the matrix, ... and, for example, the element lstlst[[2,3]] is the element in the second row and third column. In other words, in the expression lstlst[[i,j]], the first index i is the row index and the second index j is the column index. Note that the indices of elements in a list are enclosed in double square brackets.

As with simple lists, Mathematica also has a few commands to facilitate creation of special matrices. In particular, you should be aware of the statements

In[312]:= IdentityMatrix[4]
Out[312]= {{1,0,0,0},{0,1,0,0},{0,0,1,0},{0,0,0,1}}

In[313]:= DiagonalMatrix[{e,f,g,h}]
Out[313]= {{e,0,0,0},{0,f,0,0},{0,0,g,0},{0,0,0,h}}

for creating identity and diagonal matrices and the statements

In[314]:= Table[ Sin[m*x]*Sin[n*y], {m,3}, {n,3} ];
In[315]:= MatrixForm[%]
Out[315]//MatrixForm =

\[
\begin{bmatrix}
\sin(x)\sin(y) & \sin(x)\sin(2y) & \sin(x)\sin(3y) \\
\sin(2x)\sin(y) & \sin(2x)\sin(2y) & \sin(2x)\sin(3y) \\
\sin(3x)\sin(y) & \sin(3x)\sin(2y) & \sin(3x)\sin(3y) \\
\end{bmatrix}
\]

In[316]:= Quit[]

for creating and displaying a matrix whose \(mn\) element is \(\sin(mx)\sin(ny)\).

### 8.8 Using Mathematica . . .

At base, a session with Mathematica involves entering one or more expressions and then instructing Mathematica to process these expressions in some way. The following transcripts of Mathematica
sessions illustrate some of the simpler manipulations. The transcripts are presented in two columns, the one on the left recording the statements submitted to Mathematica and Mathematica’s responses, the one on the right containing explanatory comments.

Note that the manner of editing statements submitted to Mathematica depends on whether the statement is entered in the TBI or the NBI. In the TBIs on some—but not all—platforms, the arrow keys can be used to retrieve previous statements and move the cursor within a statement, the backspace and delete keys can be used to remove characters, and other keys can be used to insert characters at the position of the cursor. Further—again in some TBI’s, input lines already existing on the screen can be highlighted and then copied to the current input point.

In the NBI, in-line editing can be accomplished in the same way using the left and right arrow keys to move the cursor, the backspace or delete key to remove characters, and other keys to insert new characters. The up and down arrow keys, however, move the cursor not only through the statements on input lines but also through components of the displayed output. The cursor can be located in previous statements by typing the up arrow key a sufficient number of times (or by repositioning the cursor with a click of ML), the statement can then be edited, and the edited form reexecuted, but the new statement will take the place of the old; it will not be copied to the new input focus in the window. Alternatively, a previous statement (or portion thereof) can be highlighted, copied into the clipboard by selecting ‘Copy’ from the Edit menu, and then—after clicking ML at the desired point of insertion—selecting ‘Paste’ from the Edit menu.

Note that, whichever interface is in use, the cursor need not be positioned at the end of a line when the line is to be submitted for execution by pressing (SHIFT/ENTER); the full line will be submitted regardless of the position of the cursor within the line.

To keep the several segments in this section separated from one another, we shall periodically submit the statement Quit[]. In the TBI, this statement will terminate the Mathematica session, so we will then have to restart Mathematica; in the NBI, this statement leaves the session running in the same Notebook window. In both cases, however, the action will clear Mathematica’s workspace and, in particular, reinitialize the statement counter. The action is, of course, not necessary, and the entire “conversation” could have been “held” perfectly well without these occasional breaks or—for that matter—with breaks in different places, so long as the breaks do not occur between the point at which a variable is defined and the point at which it is used. The “conversation” in this section should be started with a fresh invocation of Mathematica.

Finally, note that, throughout this section, various components of Mathematica (functions, commands, operators, . . . ) are introduced simply by presenting statements in which they are used, frequently with brief comment. Note also that, in the NBI, various entered components will be grouped together into cells, a grouping conveyed by square brackets along the right edge of the Notebook window; the function of these groupings will be addressed briefly in Section 8.16. Details on all mentioned components and on many others can be found in the Mathematica manuals.22

8.8.1 . . . for Arithmetic

We illustrate first some of the ways in which Mathematica can be instructed to perform simple arithmetic operations. Consider, for example, the statements

\[
\begin{align*}
\text{In}[1] &= 2 + 3 \\
\text{Out}[1] &= 5
\end{align*}
\]

Enter a simple numeric expression, terminating with (SHIFT/ENTER). Mathematica returns a value, doing the arithmetic.

---

22Input statements will look much the same regardless of the platform on which we are operating. Output is here presented approximately as it would appear in the NBI. In the TBI, the content of the output will be the same but, because that interface is limited to character output, its appearance will be less “pretty”.

Enter another expression. *Mathematica* does not automatically generate a floating point evaluation but *does* reduce the fraction.

Compute $\sqrt{9}$. Note the capitalization of `Sqrt` and the *square* brackets enclosing the argument.

Compute $\sqrt{12}$. *Mathematica* simplifies the expression but keeps it exact.

Find $9! = 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$, with `Factorial`.

Find $9$ factorial with postfix operator $!$.

Ask for floating value with default number of digits.

Ask for floating value with default number of digits with postfix operator `/`. 

Ask for floating value with 20 digits.\(^{23}\)

Ask for floating value of $\pi$ with 32 digits.

\(^{23}\)While the statement \(\text{N}[4/6, 20]\) yields an evaluation with 20 digits, the statement \(\text{N}[4/6, 10]\) will on some platforms yield the same value as is displayed by the statement \(\text{N}[4/6]\). Indeed, on those platforms, \(\text{N}[4/6, n]\) yields the same value as \(\text{N}[4/6]\) for *any* $n$ less than 17. Only for $n \geq 17$ does *Mathematica* display a value with the specified number of digits. This behavior stems from the nature of the communication between the kernel (which does the computing) and the front end (which displays the results).

---

**8.8.2 ... for Algebra**

*Mathematica* is, of course, capable of much more than simple arithmetic. We next illustrate how algebraic manipulations can be requested, beginning with the statements

Enter an expression. Note that the order of the various pieces in the output may differ from that in the input. Compare the response produced on your system by this statement to the response produced by the statement \((x^2-a^2) \times (x+3)^2\).
Expand the expression. Again, the order of terms may be platform and interface dependent. Here, the symbol `%` refers to the most immediate past output. (`%%` and `%%%` refer to the penultimate and antepenultimate outputs, respectively. The symbol `%n` refers to the output on line `Out[n]`).

Factoring the expression in the previous output, assigning the result to the variable `fct`.

We note, of course, that this factoring hasn’t quite returned us to our starting point. To achieve that end, we need to multiply the first and third factors without expanding the second. To do so, we will have to carve the expression apart, manipulate with its parts, and then reconstruct the expression. As discussed in Section 8.5, Mathematica sees this expression as a structure with nested heads and lists of operands. Indeed, we can persuade Mathematica to reveal the first level in that structure by exploiting the Mathematica functions `Head` and `Part` in the statements:

Alternatively, we could exploit not only the function `Part` but also Mathematica’s ability to construct a loop in the statement

24While the number of operands is invariant, their order may be platform, interface, and perhaps context dependent. Compare the parts identified here by the integers 0, 1, 2, and 3 with the parts identified by those integers in your environment.

25We repeat the previous footnote: The order of the factors in `fct` may be platform, interface, and perhaps context dependent. Thus, the numbers identifying various operands in your environment may differ from those used here. Beware!
in which we extract parts 1, 2, 3, and 4 from \texttt{fct}, expand the product of parts 2 and 4, and then reconstruct the expression by multiplying that result by parts 1 and 3. This time, we have succeeded in reconstructing the original expression. For comparison, note that this expression has the internal representation

\begin{verbatim}
In[8]:= FullForm[fct]
Out[8]= Times[-1, Power[Plus[3, x], 2], Plus[Power[a, 2], Times[-1, Power[x, 2]]]]
\end{verbatim}

We continue with some additional manipulations. First, using the replacement operator /. and the operator \texttt{->} (which ultimately prints as \texttt{\rightarrow} in the NBI) to specify a transformation rule, we set the constant \texttt{a} to the specific value 2 with the statement

\begin{verbatim}
In[9]:= fct /. a -> 2
Out[9]= -(3 + x)^2(4 - x^2)
\end{verbatim}

Alternatively, we might have embedded the transformation rule in the command \texttt{Replace} by using the statement:\footnote{Some versions of \texttt{Mathematica} prior to Version 4.1 did not admit the third argument in the command \texttt{Replace}.}

\begin{verbatim}
In[10]:= Replace[fct, a -> 2, 3]
Out[10]= -(3 + x)^2(4 - x^2)
\end{verbatim}

Here the third argument specifies how deeply into the structure of \texttt{fct} the replacement is to be carried. Since the \texttt{a} we wish to replace appears at level 3 in \texttt{fct} (see the output of \texttt{FullForm[fct]}), we must extend the replacement to level 3 to achieve the desired substitution. However we make the replacement, we could then find a partial fraction expansion of the reciprocal of this polynomial with the statement

\begin{verbatim}
In[11]:= eq1 = Apart[1/%, x]
Find partial fraction expansion with respect to \texttt{x}.
Out[11]= 1/100(-2 + x) - 1/4(2 + x) + 1/5(3 + x)^2 + 6/25(3 + x)
\end{verbatim}

Numerous rewrites of this expression are now possible. For example, replacing \texttt{x} with \texttt{3z} would be accomplished by using the \texttt{Replace} command in the statement

\begin{verbatim}
In[12]:= Replace[eq1, x -> 3*z, 4]
Substitute \texttt{3z} for \texttt{x} in \texttt{eq1}.
Out[12]= 1/100(-2 + 3z) - 1/4(2 + 3z) + 1/5(3 + 3z)^2 + 6/25(3 + 3z)
\end{verbatim}

Alternatively, this same replacement could be accomplished with the statement

\begin{verbatim}
In[13]:= eq1 /. x -> 3*z
Substitute \texttt{3z} for \texttt{x} in \texttt{eq1}.
though we suppress the output, since it is the same as that at line Out[12].
\end{verbatim}

To illustrate a more complicated recasting, suppose that we wanted to square the denominator of the third term in \texttt{eq1} without affecting any other term. The statement \texttt{Expand[eq1]} does nothing. (Try it.) We need a more sophisticated operation that rearranges only the third term (part 3 of the original expression). To achieve the desired end, we extract that term, carve it up, manipulate with the proper piece, and substitute that piece into the proper point in the original expression. The task takes several steps, beginning with the identification of the proper part with the statements
8.8. USING MATHEMATICA ...

```
In[14]:= Part[ eq1, 3 ]
Out[14]= \frac{1}{5(x + 3)^2}

In[15]:= Part[ %, 2 ]
Out[15]= \frac{1}{(x + 3)^2}

In[16]:= Part[ eq1, 3, 2 ]
Out[16]= \frac{1}{(x + 3)^2}
```

Then we expand the denominator of this piece of `eq1` and substitute it into `eq1` as the second part of the third part with the statements

```
In[17]:= 1/Expand[1/%]
Out[17]= \frac{1}{9 + 6x + x^2}

In[18]:= ReplacePart[ eq1, %, {3, 2} ]
```

(Note that the particular part to be replaced (part `{3,2}`) is conveyed to `ReplacePart` with a list.

```
Out[18]= \frac{1}{100(-2 + x)} - \frac{1}{4(2 + x)} + \frac{6}{25(3 + x)} + \frac{1}{5(9 + 6x + x^2)}

In[19]:= Quit[]
```

We have, of course, only mentioned a few (`Expand`, `Factor`, `Part`, `Replace`, `ReplacePart`) of the numerous functions `Mathematica` makes available for manipulating expressions algebraically. We have also introduced the control structure `For`, the postfix replacement operator `/.` , and the special symbol `->` (which ultimately is displayed as `→` in the NBI). The differences among several commands, including `Expand`, `ExpandAll`, `Factor`, `Simplify`, `FullSimplify`, `Together`, `Collect`, `Apart`, `Cancel`, `Numerator`, `Denominator`, and `Factor` are explored in one of the exercises. You should make it a point to read about these functions and any associated options in the `Mathematica` manuals.

8.8.3 ... for Complex Variables

`Mathematica` is also capable of manipulating complex variables. Indeed, in most contexts, `Mathematica` assumes that any variable to which no specific value has been assigned represents a complex value. We illustrate some of `Mathematica`’s main capabilities in this category with the statements

```
In[1]:= fct = Sin[x+y*I]
Out[1]= \text{Sin}[x + i y]

In[2]:= ComplexExpand[ fct ]
Out[2]= \text{Cosh}[y] \text{Sin}[x] + i \text{Cos}[x] \text{Sinh}[y]
```

Define complex expression. The constant I stands for $\sqrt{-1}$ and is displayed as `i` in output.

Recast `fct` in form $a + ib$. In contrast to the default behavior, the function `ComplexExpand` assumes that variables to which no value has been assigned represent real values.
In[3]:= ComplexExpand[ Re[fct] ]

In[4]:= ComplexExpand[ Im[fct] ]

In[5]:= ComplexExpand[ Conjugate[fct] ]

Extract real part of fct.
Extract imaginary part of fct.
Find complex conjugate of fct.

In[6]:= fct1 = (x+i*y)/(x-i*y)
Out[6]= x + i y

In[7]:= ComplexExpand[ fct1, TargetFunctions -> {Re, Im} ]
Out[7]= x^2/\(x^2 + y^2\) + 2 i x y/\(x^2 + y^2\) - y^2/\(x^2 + y^2\)

Recast fct1 in form a + ib. ComplexExpand usually tries to express its output in terms of the real and imaginary parts of the input variables. Sometimes, however, the option TargetFunctions must be used to specify which of Re, Im, Abs, Arg, Conjugate and Sign are to be used. We include that option here simply to illustrate it; the option is not actually necessary in this context.

Find complex absolute value of previous result.
And simplify it.

In[8]:= ComplexExpand[ Abs[fct1], TargetFunctions -> {Re, Im} ]
Out[8]= \sqrt{\frac{4 x^2 y^2}{(x^2 + y^2)^2} + \left(\frac{x^2}{x^2 + y^2} - \frac{y^2}{x^2 + y^2}\right)^2}

Find complex absolute value of fct1, noting that this route includes the simplification that we had to request explicitly in the previous route.

In[9]:= Simplify[ % ]
Out[9]= 1

In[10]:= ComplexExpand[ Abs[fct1] ]
Out[10]= 1

Find complex argument of fct1. In Mathematica, ArcTan[a, b] = tan⁻¹(b/a).

To conclude this section, we record quickly how to convert complex numbers from Cartesian form \(z = x + iy\) to polar form \(z = |z|e^{i\text{Arg}(z)}\) and from polar form to Cartesian form. Starting with a complex number in Cartesian form, we would use statements like

In[13]:= fct3 = x + I*y
Out[13]= x + iy

In[14]:= ComplexExpand[ Abs[fct3], TargetFunctions -> {Re, Im} ]
Out[14]= \sqrt{x^2 + y^2}

Find complex absolute value of fct1.

In[15]:= ComplexExpand[ Arg[fct3], TargetFunctions -> {Re, Im} ]
Out[15]= ArcTan\left[\frac{x^2 - y^2}{x^2 + y^2}, \frac{2 y x}{x^2 + y^2}\right]

Simplify it, though Mathematica doesn’t seem to recognize that \(\text{ArcTan}[x^2 - y^2, 2 y x]\) would be even simpler.

In conclusion, we record quickly how to convert complex numbers from Cartesian form \(z = x + iy\) to polar form \(z = |z|e^{i\text{Arg}(z)}\) and from polar form to Cartesian form. Starting with a complex number in Cartesian form, we would use statements like

In[13]:= fct3 = x + I*y
Out[13]= x + iy

In[14]:= ComplexExpand[ Abs[fct3], TargetFunctions -> {Re, Im} ]
Out[14]= \sqrt{x^2 + y^2}

In[15]:= ComplexExpand[ Arg[fct3], TargetFunctions -> {Re, Im} ]

to find the absolute value and argument of the number. Conversely, we would use statements like
8.8. USING MATHEMATICA ...

In[16]:= fct4 = r * Exp[I*\[Theta] ]
Out[16]= \[E]^{i \[Theta]} \[R]r
In[17]:= ComplexExpand[ ExpToTrig[fct4] ]
In[18]:= Quit[]

to find the Cartesian expression of a complex number originally given in polar form.

We have, of course, only mentioned some (ComplexExpand, Re, Im, Conjugate, Abs, Arg, ExpToTrig) of the functions Mathematica makes available for manipulating complex numbers. Further, we have also introduced the option TargetFunctions for the command ComplexExpand, and we have seen additional contexts in which the command Simplify can be valuable. You should make it a point to read about these functions and any associated options in the Mathematica manuals.

8.8.4 ... for Trigonometry

Mathematica also makes available a number of commands and functions for manipulating with trigonometric, hyperbolic, and exponential functions. Note, for example, the capabilities illustrated in the statements

In[1]:= trig1 = Sin[x] * Cos[x]
In[2]:= TrigToExp[ trig1 ]
In[3]:= ExpToTrig[ % ]
In[4]:= TrigReduce[ trig1 ]
In[5]:= TrigExpand[ % ]
In[6]:= trig2 = Sin[2*x+y]
In[7]:= TrigExpand[ trig2 ]
- \[Sin]x^2 \[Sin]y
In[8]:= TrigReduce[ % ]
Out[8]= \[Sin]2x+y
In[9]:= TrigExpand[ trig2/.x->z/2/.z->2*x ]
In[10]:= TrigExpand[ % ]
- \[Sin]x^2 \[Sin]y

Define a trig expression.
Convert to exponential form.
Convert to standard trig functions.
Recast with multiple angle identities.
Expand functions of multiple angles.
Define another trig function.
Expand fully using trig addition formulae.
Return to original form.
Expand, but don't expand 2x. Here, we replace 2x with z for the expansion but then replace z with 2x after the expansion has been evaluated.
Expand again to complete expansion.
In[11]:= (Cos[x] + Sin[x])^2
Out[11]= (Cos[x] + Sin[x])^2

In[12]:= Expand[ % ]
Out[12]= Cos[x]^2 + 2 Cos[x] Sin[x] + Sin[x]^2

In[13]:= Quit[]

We have, of course, only mentioned a few (TrigToExp, ExpToTrig, TrigReduce, TrigExpand, and Expand) of the functions Mathematica makes available for doing trigonometry. You should make it a point to read about these and other commands and any applicable special arguments or options in the Mathematica manuals.

8.8.5 ... for Algebraic Equations

We turn next to illustrating a few Mathematica commands for manipulating expressions algebraically. To define a polynomial and find its roots, for example, we would use statements like

\[
\text{In[1]:= poly = Expand[ (x^2-a^2)*(x+3)^2 ];}
\]

Define a polynomial.

\[
\text{In[2]:= soln = Solve[ {poly == 0}, {x} ];}
\]

Solve the equation \( poly = 0 \) for \( x \). There are four roots, though two of them are identical.

\[
\text{Out[2]= \{ \{x \rightarrow -3\}, \{x \rightarrow -3\}, \{x \rightarrow -a\}, \{x \rightarrow a\} \}}
\]

Note that the arguments to the command \text{Solve} are lists, and that the roots have been returned as a list of lists of transformation rules—a behavior that would seem more reasonable had we been solving a system of equations for more than one unknown. In the simple case of a single equation and a single unknown, the braces can be omitted, as in

\[
\text{In[3]:= soln1 = Solve[ poly == 0, x ];}
\]

\[
\text{Out[3]= \{ \{x \rightarrow -3\}, \{x \rightarrow -3\}, \{x \rightarrow -a\}, \{x \rightarrow a\} \}}
\]

Verify first root.

Verify all roots in a single statement.

Verify all roots, using the \text{For} structure to code a loop.

Perhaps the most famous of all solutions is that of a quadratic equation. The command \text{Solve} clearly knows the quadratic formula, as illustrated by the statements
8.8. USING MATHEMATICA ...

Supply a quadratic expression.

\[ ax^2 + bx + c = 0 \] for \( x \). Note again that the roots have been returned as a list of lists.

Evaluate solution with indicated values of \( a, b, \) and \( c \).

We can, of course, always ask for floating point evaluations with the command \( \text{N} \), finding here that

We have accepted the default value (6) for the number of digits in the floating point values returned by \( \text{N} \).

Finally, let us illustrate \textit{Mathematica}'s capability to solve simultaneous linear equations. To define and solve a representative system, we invoke the statements

\[
\begin{align*}
\text{In}[11]:= & \text{eqn1 = } x - y == 4 \\
\text{Out}[11]= & x - y == 4 \\
\text{In}[12]:= & \text{eqn2 = } x + y == 0 \\
\text{Out}[12]= & x + y == 0 \\
\text{In}[13]:= \text{Solve[} \{\text{eqn1,eqn2}\},\{x,y\} ] \\
\text{Out}[13]= & \{x \to 2, y \to -2\}
\end{align*}
\]

Note that the equations and the variables are presented to \text{Solve} as lists and that each solution—here there is only one—is presented as a list. Were there several solutions, \textit{Mathematica} would have returned a list of lists.

We need not, of course, have specific numerical values for the coefficients in the equations to be solved. Solution of a more symbolic system of linear equations is illustrated in the statements

\[
\begin{align*}
\text{In}[14]:= & \text{eqn1 = } a*x + b*y == e; \\
\text{In}[15]:= & \text{eqn2 = } c*x + d*y == f; \\
\text{In}[16]:= \text{Solve[} \{\text{eqn1,eqn2}\},\{x,y\} ] \\
\text{Out}[16]= & \{x \to \frac{-c e + a f}{b c - a d}, y \to \frac{-d e - b f}{b c - a d}\}
\end{align*}
\]

Remove variables \textit{eqn1} and \textit{eqn2} from \textit{Mathematica}'s memory.

We have, of course, only mentioned one (\text{Solve}) of the functions \textit{Mathematica} makes available for solving algebraic equations exactly. Your attention is drawn also to the function \text{Nsolve} for finding solutions numerically and to the function \text{LinearSolve}, which is built to solve matrix equations of the form \( Ax = b \). You should make it a point to read about these functions in the \textit{Mathematica} manuals. Rootfinding is discussed in greater depth in Chapter 14.
8.8.6 ... for Manipulating Vectors and Matrices

Vectors and matrices represent numerous physical quantities. We have already seen in Section 8.7 how to create lists. Vectors are simply lists and (two-dimensional) matrices are lists of lists. Once they have been created, Mathematica can manipulate these objects in numerous ways. To illustrate some of these capabilities, we begin by defining a matrix and evaluating its transpose with the statements

\begin{verbatim}
In[1]:= mat1 = { {1,2}, {3,4} };
In[2]:= MatrixForm[ mat1 ]
Out[2]//MatrixForm=
\begin{pmatrix}
1 & 2 \\
3 & 4 \\
\end{pmatrix}

In[3]:= mat2 = Transpose[ mat1 ];
In[4]:= MatrixForm[ mat2 ]
Out[4]//MatrixForm=
\begin{pmatrix}
1 & 3 \\
2 & 4 \\
\end{pmatrix}
\end{verbatim}

Then, we evaluate the inverse, the trace, and the determinant of the matrix with the statements

\begin{verbatim}
In[5]:= MatrixForm[ Inverse[ mat1 ] ]
Out[5]//MatrixForm=
\begin{pmatrix}
-2 & 1 \\
3 & -1 \\
2 & -2 \\
\end{pmatrix}

In[6]:= val1 = Tr[ mat1 ]

In[7]:= val2 = Det[ mat1 ]
\end{verbatim}

Other common manipulations include evaluating the standard matrix product and an element-by-element product. The standard (non-commutative) matrix product is specified by the infix operator ‘.’ in a statement like

\begin{verbatim}
In[8]:= MatrixForm[ mat1.mat2 ]
Out[8]//MatrixForm=
\begin{pmatrix}
5 & 11 \\
11 & 25 \\
\end{pmatrix}
\end{verbatim}

Element-by-element (commutative) multiplication is accomplished with the infix operator ‘∗’ in a statement like

\begin{verbatim}
In[9]:= MatrixForm[ mat1*mat2 ]
Out[9]=
\begin{pmatrix}
1 & 6 \\
6 & 16 \\
\end{pmatrix}
\end{verbatim}

Finally (for this section anyway), Mathematica includes functions for finding eigenvalues and eigenvectors, which we illustrate with the statements
In[10]:= spinx = {{0, 1/Sqrt[2], 0}, {1/Sqrt[2], 0, 1/Sqrt[2]}, {0, 1/Sqrt[2], 0}};

Enter a matrix bound to the variable spinx.

In[11]:= MatrixForm[ % ]

Display the matrix in matrix form.

Out[11]=
\[
\begin{pmatrix}
0 & 1 & 0 \\
\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\
0 & \frac{1}{\sqrt{2}} & 0
\end{pmatrix}
\]

In[12]:= Eigenvalues[ spinx ]

Find eigenvalues.

Out[12]= \{-1, 1, 0\}

In[13]:= Eigenvectors[ spinx ]

Find eigenvectors.

Out[13]= \{
\{1, -\sqrt{2}, 1\}, \{1, \sqrt{2}, 1\}\{-1, 0, 1\}\}

In[14]:= Eigensystem[ spinx ]

Find eigenvalues and eigenvectors with a single statement.

Out[14]= \{
\{-1, 0, 1\}, \{1, -\sqrt{2}, 1\}, \{-1, 0, 1\}, \{1, \sqrt{2}, 1\}\}\n
In[15]:= MatrixForm[%]

Display result in matrix form, in which the first row contains the eigenvalues and the second row contains the associated eigenvectors.

Out[15]=
\[
\begin{pmatrix}
-1 & 1 & 0 \\
1, -\sqrt{2}, 1 & 1, \sqrt{2}, 1 & -1, 0, 1
\end{pmatrix}
\]

In[16]:= Quit[]

The output of Eigenvalues is a list of the eigenvalues of the specified matrix. The output of Eigenvectors is a list of lists, each of the latter of which is the (unnormalized) eigenvector belonging to the eigenvalue in the same position in the list returned by Eigenvalues. Finally, the output of Eigensystem is a two-element list, the first element of which is the output of Eigenvalues and the second element of which is the output of Eigenvectors.

We have, of course, only mentioned a few (MatrixForm, Transpose, Inverse, Tr. Det, Eigenvalues, Eigenvectors, Eigensystem) of the functions available for manipulating matrices. Important additional functions include AppendColumns, AppendRows, BlockMatrix, TakeRows, TakeColumns, TakeMatrix, SubMatrix, ZeroMatrix, and LinearEquationsToMatrix, which are contained in the add-on package LinearAlgebra'MatrixManipulation and the function Normalize, which is contained in the add-on package LinearAlgebra'Orthogonalization. You should make it a point to read about these functions and any associated arguments in the Mathematica manuals.

8.8.7 ... for Basic Calculus

Differentiation, indefinite and definite integration, and determination of series representations of symbolic expressions are also among Mathematica’s capabilities. To evaluate derivatives, for example, we would use statements like

\[\text{27 Packages are discussed in Section 8.9.}\]
Enter an expression.

Differentiate it with respect to $x$.

Differentiate it twice with respect to $x$ and simplify the result.

To evaluate integrals, we would instead use statements like

Evaluate indefinite integral of the first derivative.

Enter another expression.

Evaluate definite integral on $x$ from $x=0$ to $x=+\infty$.

Interestingly, Mathematica in this case returns a result revealing that the integral is not well defined. In the returned conditional expression the first argument is the evaluation of the integral, but only when the condition in the second argument is true. If the specified condition is false ($\text{Re}(a)=0$), then the denominator in the integrand passes through zero somewhere in the interval of integration and the integrand is indeed ill defined. If, however, the specified condition is true, then the integrand is well defined, having the value $1/\sqrt{a^2}$, a result not written as $1/a$ because $a$ is not known to be positive or even real.

We can instruct Mathematica to be more specific by exploiting an optional argument to `Integrate`, as for example in the statements

Tell Mathematica that $a$ is real.

Tell Mathematica that $a$ is positive, which implies as well that $a$ must be real.

Finally (for this section), we point out that Mathematica can generate series expansions of prescribed functions:

Enter yet another expression.

Expand the expression in a Taylor series in $x$ about $x=0$, retaining terms through those of 7-th order in $x$. Note here that the coefficients of the terms in $x^2$, $x^4$, and $x^6$ are all zero.

While a series displays on the screen as an algebraic expression, internally its form differs from that of many other (i.e., of ordinary) expressions. The statement
8.8. USING MATHEMATICA ...

In[11]:= FullForm[ ser ]
Out[11]/FullForm= SeriesData[ x, 0, List[ Rational[1,2], Rational[1,4], 0, Rational[-1,48],
        0, Rational[1,480], 0, Rational[-17,80640] ], 0, 8, 1 ]

reveals that the head of an expression that conveys a series is SeriesData. The remaining parts of
the expression are, in order, the variable \( x \) in which the series is expressed, the value \( x_0 \) (here 0) of
\( x \) about which the expansion is made, a list of the coefficients of the successive powers of \( x - x_0 \) in
the expansion, and two arguments (here 0 and 8) which convey the starting and ending powers as
multiples of the final argument (here 1). The powers appearing thus are 0/1, 2/1, 3/1, ... 8/1.

Unfortunately, for some operations, the expression of a series in the form dictated by SeriesData
cannot be used. In those cases, we may need to convert the series into an ordinary expression by
invoking the command Normal with the statement

In[12]:= pol = Normal[ ser ]
Out[12]= 1
Out[12]= \frac{1}{2} + \frac{x}{4} + \frac{x^3}{48} + \frac{x^5}{480} - \frac{17x^7}{80640}

In[13]:= FullForm[ pol ]
Out[13]= Plus[ Rational[1,2], Times[ Rational[1,4], x ],
    Times[ Rational[-1,48], x^3 ],
    Times[ Rational[1,480], x^5 ],
    Times[ Rational[-17,80640], x^7 ] ]

In[14]:= Quit[]

Note that the omitted terms in the series are not indicated in \( \text{pol} \) and that \( \text{pol} \) has expression type
Plus. Further, compare the storage of a series as revealed by the statement FullForm[ser] with
the storage of a polynomial as revealed by the statement FullForm[pol].

By default, Mathematica commands automatically carry out particular operations that, in
Mathematica’s mind, simplify expressions submitted to them. As we have seen above, the command D for differentiation, for example, will result in immediate evaluation of any derivatives that
can be evaluated. Sometimes, however, we may wish to enter an expression and suppress immedi-
ate evaluation. Mathematica provides the command HoldForm for that purpose and the command
ReleaseHold to request explicit evaluation at a later time. To illustrate, suppose we define the
function

In[1]:= f[y_] := a * Sin[ \[Omega] * y ]

Then, the two statements

In[2]:= D[ f[y], y ]
Out[2]= a \omega \cos[ y \omega ]

In[3]:= d1 = HoldForm[ D[ f[y], y ] ]
Out[3]= \partial_y f[y]

reveal the difference. In the first, the derivative is evaluated; in the second, it is left unevaluated.
At some later time, we could evaluate the derivative in this last expression with the statement

In[47]:= ReleaseHold[ d1 ]
Out[47]= a \omega \cos[ y \omega ]
This feature might, for example, be useful if we had several functions whose derivative we wished to evaluate. We would simply define the derivative with \texttt{HoldForm}. Then, we could evaluate the derivative of each new function by changing the definition of \( f[y] \) before executing the simple statement illustrated on Line In[47].

\texttt{In[48]}:= \texttt{Quit[]}

We have, of course, only mentioned a few (\texttt{D, \texttt{Integrate}, \texttt{Series}, \texttt{Normal}) of the numerous functions \texttt{Mathematica} makes available for doing calculus. One important additional function is \texttt{Limit}. You should make it a point to read about these functions and any associated options in the \texttt{Mathematica} manuals.

8.8.8 ... for Laplace Transforms

In Section 1.5.2, we defined the Laplace transform \( \tilde{f}(s) \) of a function \( f(t) \) by the integral

\[
\tilde{f}(s) = \int_0^\infty e^{-st} f(t) \, dt
\]

(8.1)

and showed a number of the properties exhibited by this integral transform. Among other properties, the Laplace transform has the capacity to convert a linear, ordinary differential equation into a linear algebraic equation for the Laplace transform of the solution. Thus, as we shall see in greater detail in Chapter 11, one effective strategy for solving a linear ordinary differential equation is to take its Laplace transform, solve the resulting algebraic equation for that transform, and then invert that transform to return from the \( s \)-space of the transform to the \( t \)-space (time-space) of the original problem. The initial conditions are automatically incorporated in this approach. Note, unfortunately, that inverting the Laplace transform of the solution to find the solution itself is rarely easy.

\texttt{Mathematica} provides for quick evaluation of an assortment of integral transforms. In particular the function \texttt{LaplaceTransform} will calculate a Laplace transform and the function \texttt{InverseLaplaceTransform} will invert a transform to return to the original space. We illustrate the use of those functions with the statements

\texttt{In[1]}:= \texttt{Sin[\Omega \ast t];}

\texttt{In[2]}:= \texttt{LaplaceTransform[ \%, \textit{t}, \textit{s} ]}

\texttt{Out[2]}= \frac{\omega}{s^2 + \omega^2}

\texttt{In[3]}:= \texttt{InverseLaplaceTransform[ \%, \textit{s}, \textit{t} ]}

\texttt{Out[3]}= \texttt{Sin[t \omega]}

\texttt{In[4]}:= \texttt{LaplaceTransform[ \texttt{Exp[-a \ast t]}, \textit{t}, \textit{s} ]}

\texttt{Out[4]}= \frac{1}{a + s}

\texttt{In[5]}:= \texttt{Simplify[ InverseLaplaceTransform[}

\texttt{1/( (s+a)*(s+b) ), \textit{s}, \textit{t} ] ]}

\texttt{Out[5]}= - \frac{e^{-at} - e^{-bt}}{a - b}

\texttt{In[6]}:= \texttt{Quit[]}

Define a function. Note that the expression \( \Omega \) will ultimately print as \( \omega \).

Calculate the Laplace transform of the function, where \( t \) is the primary variable and \( s \) is the variable in terms of which the transform is to be expressed.

Then, invert the transform to return to the original function.

Calculate the Laplace transform of the function \( e^{-at} \).

Calculate and simplify the inverse Laplace transform of the function \((s+a)^{-1}(s+b)^{-1}\).
8.8. USING MATHEMATICA . . .

8.8.9 . . . for Ordinary Differential Equations

Among Mathematica’s strongest suits is its ability to solve ordinary differential equations (ODEs). We will return in Chapter 11 to a more extended discussion of those capabilities. For this chapter, we limit ourselves to a quick illustration, choosing the motion of an object of mass \( m \) attached to a spring having constant \( k \) and to a dashpot (shock absorber) having damping constant \( b \). The appropriate equation of motion is communicated to Mathematica with the statement\(^{28}\)

\[
\text{In}[1]:= \text{deq1} = m\cdot x''[t] + b\cdot x'[t] + k\cdot x[t] == 0
\]

\[
\text{Out}[1]= k\cdot x[t] + b\cdot x'[t] + m\cdot x''[t] == 0
\]

To solve this equation for \( x \) as a function of \( t \), we can proceed in either of two ways. If we submit the statement

\[
\text{In}[2]:= \text{soln1} = \text{DSolve}[\text{deq1}, x[t], t]
\]

Mathematica returns the solution in the form\(^{29}\)

\[
\text{Out}[2]= \begin{cases} 
  x[t] \to e^{(-b-\sqrt{b^2-4km})t/2m} C[1] + e^{(-b+\sqrt{b^2-4km})t/2m} C[2] 
\end{cases}
\]

Here, \text{DSolve} has used the symbols \( C[1] \) and \( C[2] \) for the initially undetermined integration constants in the solution to a second-order differential equation. Alternatively, we might phrase the invocation of \text{DSolve} in the form

\[
\text{In}[3]:= \text{soln2} = \text{DSolve}[\text{deq1}, x, t]
\]

in which case \text{Mathematica} returns the solution in the form

\[
\text{Out}[3]= \begin{cases} 
  x \to \text{Function}[\{t\},e^{(-b-\sqrt{b^2-4km})t/2m} C[1] + e^{(-b+\sqrt{b^2-4km})t/2m} C[2]]
\end{cases}
\]

where again the symbols \( C[1] \) and \( C[2] \) have been used for the integration constants.

The two forms of the solution look similar. Each is expressed as a transformation rule. The solution \( \text{soln1} \), however, is not a function. The rule expressing \( \text{soln1} \) allows substitution of the solution for \( x[t] \) in any expression. For example, the statement

\[
\text{In}[4]:= x[t] + \text{D}[x[t], t] /. \text{soln1}
\]

yields a result in which \( x[t] \) has been replaced in the first term but the derivative has been left unevaluated. At times, this first form of the solution will be adequate. If, however, we expect to manipulate with the solution, the second form is preferable, since it expresses the solution as a true function of the independent variable \( t \). Had we used \( \text{soln2} \) in the example at \text{In}[4], i.e., had we submitted the statement

\[
\text{In}[5]:= x[t] + \text{D}[x[t], t] /. \text{soln2}
\]

\(^{28}\)When the variable with respect to which a derivative is to be evaluated is the only quantity on which a function depends, we can use a prime ‘ to convey differentiation rather than the more explicit form \( \text{D}[x[t], t] \).

\(^{29}\)As is often the case, we here present the output in a form that is more transparent than—i.e., different from—the form presented directly in the Mathematica output. In what follows, if we extract explicit parts of the expressions, we will do so in ways that do not depend on these differences.
Mathematica would have returned a more complicated result in which the solution would have been substituted both in the first and in the second term and then the derivative would have been evaluated.

Whichever form of the solution we have requested, Mathematica’s response to the simple statements we have used has been blindly to write a solution that is mathematically correct but has not recognized that the most convenient form of this solution will depend on the algebraic sign of the quantity under the square root in the two exponents. There are three cases:

1. $b^2 > 4km$ (overdamped), in which case the square roots are real and both terms have real—though in both cases decaying—exponentials.

2. $b^2 = 4km$ (critically damped), in which case the two terms are the same and Mathematica has returned only one solution when the general solution should be a linear superposition of two linearly independent solutions. The returned solution is mathematically correct, but we must carefully evaluate the limit as $b^2$ approaches $4km$ to extract a more transparent form.

3. $b^2 < 4km$ (underdamped), in which case the square roots are imaginary and the complex exponentials that then appear would more appropriately be written in terms of sines and cosines of real arguments, each multiplied by the decaying exponential $e^{-bt/2m}$.

To achieve a more appropriate form in the case of underdamped motion, we need to tell Mathematica that $b^2 < 4km$, an end that is (unfortunately) not easily accomplished, partly because Mathematica assumes that variables to which no value has been assigned represent complex quantities and does not have a facility for overriding that assumption except within the context of a few specific commands. The following sequence of statements will achieve the desired objective.\(^{30}\) We begin by using the statement

\begin{verbatim}
In[6]:= soln3 = soln1[[1]][[1]][[2]]
Out[6]= e^{(-b-\sqrt{b^2-4km})t/2m} C[1] + e^{(-b+\sqrt{b^2-4km})t/2m} C[2]
\end{verbatim}

to extract the second part of the first part of the first part of soln1, i.e., to extract an expression giving the solution itself. Then, we simplify the appearance of this expression by introducing (temporarily) the variables

$$\gamma = \frac{b}{2m} \quad \text{and} \quad \omega = \sqrt{\frac{k}{m}} \quad (8.2)$$

with the statement

\begin{verbatim}
In[7]:= soln4 = soln3 /. { b -> 2*\[Gamma]*m, k -> \[Omega]^2*m }
Out[7]= e^{t(-2m\gamma-\sqrt{4m^2\gamma^2-4m^2\omega^2})/2m} C[1] + e^{t(-2m\gamma+\sqrt{4m^2\gamma^2-4m^2\omega^2})/2m} C[2]
\end{verbatim}

Next, we invoke the statement\(^{31}\)

\begin{verbatim}
In[8]:= soln5 = Simplify[ soln4, m > 0 ]
Out[8]= e^{-t(\gamma+\sqrt{\gamma^2-\omega^2})} \left( C[1] + e^{2t(\gamma+\sqrt{\gamma^2-\omega^2})} C[2] \right)
\end{verbatim}

to clear the $m$’s, the statement

\begin{verbatim}
In[9]:= soln6 = Expand[ soln5 ]
Out[9]= e^{-t(\gamma+\sqrt{\gamma^2-\omega^2})} C[1] + e^{2t(\gamma+\sqrt{\gamma^2-\omega^2})-t(\gamma+\sqrt{\gamma^2-\omega^2})} C[2]
\end{verbatim}

\(^{30}\)The author continues to search for less cumbersome routes but, so far anyway, has failed to find any.

\(^{31}\)Some versions of Mathematica prior to Version 4.1 did not admit the second argument in the command Simplify.
to distribute the initial factor over both terms, and the command Map in the statement

\[ \text{In}[10]:= \text{soln7} = \text{Map}[\text{Expand}, \text{soln6}, 3] \]

\[ \text{Out}[10]= e^{-t\gamma - \sqrt{\gamma^2 - \omega^2}} C[1] + e^{-t\gamma + \sqrt{\gamma^2 - \omega^2}} C[2] \]

to apply Expand to some of the internal pieces of the expression in soln6. Finally, we simplify the exponents further by recognizing that the square roots are, in fact imaginary (when \( 4km > b^2 \), i.e., \( \omega > \gamma \)), a recognition achieved by making the substitution

\[ \text{In}[11]:= \text{soln8} = \text{soln7} /. \text{Sqrt}[\gamma^2 - \omega^2] \rightarrow i*a \]

\[ \text{Out}[11]= e^{-iat-t\gamma} C[1] + e^{iat-t\gamma} C[2] \]

At last, we are closing in on a more convenient presentation of the solution for the underdamped motion of a harmonic oscillation. In essence, we wish to replace the exponentials \( e^{\pm iat} \) with \( \cos at \pm i\sin at \). We take the next step with the statement

\[ \text{In}[12]:= \text{soln9} = \text{Factor}[\text{Map}[\text{ComplexExpand}, \text{soln8}]] \]

\[ \text{Out}[12]= e^{-t\gamma} \left( C[1] \cos at + C[2] \cos at - iC[1] \sin at + iC[2] \sin at \right) \]

From this solution, however, we note that the expression would have a yet simpler form if we substituted \( A \) for \( C[1] + C[2] \) and \( B \) for \(-i(C[1] - C[2])\). We achieve that objective with the statements

\[ \text{In}[13]:= \text{Solve}\{C[1]+C[2]==A, -i(C[1]-C[2])==B\}, \{C[1],C[2]\}] \]

\[ \text{Out}[13]= \left\{ \left\{ C[1] \rightarrow \frac{1}{2} (A + iB), C[2] \rightarrow \frac{1}{2} (A - iB) \right\} \right\} \]

\[ \text{In}[14]:= \text{soln10} = \text{Simplify}[\text{soln9} /. \%[[1]]] \]

\[ \text{Out}[14]= e^{-t\gamma} \left( A \cos at + B \sin at \right) \]

We, of course, are not finished until we have imposed suitable initial conditions on the solution. To do so, we determine \( A \) and \( B \) to reflect the simple assumption that the motion is started by drawing the oscillator aside and releasing it from rest, i.e., that

\[ x(0) = x_0 ; \quad \frac{dx}{dt}(0) = 0 \quad (8.3) \]

Then, we find the constraints imposed on the integration constants with the statements

\[ \text{In}[15]:= \text{eq1} = x_0 == \text{soln10} /. t \rightarrow 0 \]

\[ \text{Out}[15]= x_0 == A \]

\[ \text{In}[16]:= \text{eq2} = 0 == D[\text{soln10}, t] /. t \rightarrow 0 \]

\[ \text{Out}[16]= 0 = aB - A\gamma \]

\[ \text{In}[17]:= \text{Solve}\{\text{eq1,eq2}, \{A,B\}\} \]

\[ \text{Out}[17]= \left\{ \left\{ B \rightarrow \frac{x_0 \gamma}{a}, A \rightarrow x_0 \right\} \right\} \]

\[ \text{In}[18]:= \text{soln11} = \text{soln10} /. \%[[1]] \]

\[ \text{Out}[18]= e^{-t\gamma} \left( x_0 \cos at + \frac{x_0 \gamma \sin [at]}{a} \right) \]

\[ \text{In}[19]:= \text{soln12} = \text{Collect}[\text{soln11}, x_0] \]

\[ \text{Out}[19]= e^{-t\gamma} x_0 \left( \cos at + \frac{\gamma \sin [at]}{a} \right) \]
To interpret this result in terms of the original dimensional parameters, we need merely remember that

\[ a = \sqrt{\omega^2 - \gamma^2}; \quad \gamma = \frac{b}{2m}; \quad \omega^2 = \frac{k}{m} \]  

(8.4)

These values could, of course, be reinserted into the solution obtained at line Out[19]. At long last, we have reached a form of the solution that reveals its character as an oscillation with an exponential decay of its amplitude. We leave it to the reader to use Mathematica to verify that this solution in fact satisfies the original equation and the imposed initial conditions. (Similar simpler expressions for the critically and overdamped cases are deduced in the exercises.)

In the above example, we chose to leave the initial conditions unspecified until after the solution had been obtained and recast in a more transparent form. Alternatively, we could have incorporated the initial conditions in the original invocation of \texttt{DSolve} with an expanded statement of the form

\( \text{In}[20]:= \text{soln13} = \text{DSolve} \{ \text{deq1, x[0]==x0, x'[0]==0 }, x[t], t \} \)

The resulting solution would still have been presented in an exponential form in which, for the underdamped case, the square roots appearing would have been imaginary. We would still have to work our way through a number of transformations before having the solution in the standard form for the underdamped case. Note specifically that the equations and initial conditions have here been presented to \texttt{DSolve} as a list.

Much of our manipulation in the above illustration would have been simplified if we had expressed the original equation in dimensionless form before seeking its solution. To that end, we submit the statements

\( \text{In}[21]:= \text{Quit}[] \)
\( \text{In}[1]:= \text{deq1} = m*x''[t] + b*x'[t]+k*x[t]==0 \)
\( \text{Out}[1]= kx[t] + bx'[t] + mx''[t] == 0 \)

to clear the workspace and redefine the equation. Then, with the statements

\( \text{In}[2]:= \text{deq2} = \text{deq1} /. \{ \text{k} \to m*\Omega^2, \text{b} \to 2*m*\gamma \} \)
\( \text{Out}[2]= m\omega^2 x[t] + 2m\gamma x'[t] + mx''[t] == 0 \)
\( \text{In}[3]:= \text{deq3} = \text{Cancel} \{ \text{deq2[[1]]/m } \} == 0 \)
\( \text{Out}[3]= \omega^2 x[t] + 2\gamma x'[t] + x''[t] == 0 \)

we introduce the quantities \( \omega \) and \( \gamma \) defined by \( \omega^2 = k/m \) and \( \gamma = b/(2m) \), where \( \omega \) is the natural frequency of the oscillator, and remove the parameter \( m \) from the picture altogether. Next, we change the independent variable to \( \tau = \omega t \). One successful process for achieving that end involves

(1) introducing a new function with the statement

\( \text{In}[4]:= \text{x[t_,} := y[\Omega*t] \)

(2) reevaluating \texttt{deq3} with the statement

\( \text{In}[5]:= \text{deq4} = \text{deq3} \)
\( \text{Out}[5]= \omega^2 y[t \omega] + 2\gamma y'[t \omega] + \omega^2 y''[t \omega] == 0 \)

(3) replacing the quantity \( t \omega \) (which is now the variable with respect to which differentiation is to be evaluated in \texttt{deq4}) with a single variable \( \tau \) by executing the statement
In[6]:= deq5 = deq4 /. t*\[Omega] -> \[Tau]

(4) introducing a new dimensionless variable \( \beta \) defined by \( \gamma = \omega \beta \) with the statement

In[7]:= deq6 = deq5 /. \[Gamma] -> \[Omega]*\[Beta]

and, finally, (5) removing the common factor of \( \omega^2 \) with the statement

In[8]:= deq7 = Expand[ deq6[[1]]/\[Omega]^2 ] == 0

At this point, we have swallowed all of the dimensional constants either into a rescaling of the time (\( \tau = \omega t \)) or into a single dimensionless parameter (\( \beta = b/m\omega \)). We could, of course, rescale what was originally \( x(t) \) and is now \( y(\tau) \) to some chosen dimensionless quantity, but the equation is linear in this quantity and we can simply interpret the equation as it stands in those terms. We would then solve the equation as before with the statement

In[9]:= soln = DSolve[ {deq7, y[0]==x0, y'[0]==0}, y[\[Tau]], \[Tau] ]
Out[9]= \{ \{ y[\tau] \to \frac{1}{2\sqrt{-1+\beta^2}} \left[ x_0 \left( -e^{-(-\beta-\sqrt{-1+\beta^2})\tau} \beta + e^{-(-\beta+\sqrt{-1+\beta^2})\tau} \beta + e^{-(-\beta+\sqrt{-1+\beta^2})\tau} \sqrt{-1+\beta^2} + e^{-(-\beta-\sqrt{-1+\beta^2})\tau} \sqrt{-1+\beta^2} \right) \} \}

We leave it to the reader to use Mathematica to show that this solution is equivalent to the result shown at line Out[19] when \( \beta < 1 \).\(^{32}\)

In[10]:= Quit[]

We have, of course, mentioned only the command DSolve as a function for solving ordinary differential equations. We have also seen additional contexts in which replacement of variables with the operator ‘/.’ and the commands Expand, Factor, Solve, ComplexExpand, and Simplify are useful. Finally, we have met the commands Map, Collect, and Cancel for the first time. You should make it a point to read about these commands and any associated options in the Mathematica manuals.

8.8.10 ... for Vector Calculus

Mathematica includes several commands to facilitate vector calculus.\(^{33}\) The main commands are Grad, Div, Curl, and Laplacian for the gradient, divergence, curl, and Laplacian, respectively. Each accepts a first argument—a scalar function for Grad and Laplacian; a vector function for Div and Curl—and a second argument, a list that identifies the variables in the coordinate system in which the quantity is expressed. An optional third argument—a string—specifies the coordinate system to be used, with "Cylindrical" and "Spherical" as the main alternatives to the default ("Cartesian").

\(^{32}\)Here, the assumption \( \beta < 1 \) is equivalent to the assumption \( \gamma/\omega < 1 \) or \( \gamma^2 < \omega^2 \) or \( b^2/4m\omega^2 < k/m \) or \( b^2 < 4km \).

\(^{33}\)Prior to Version 9, vector calculus was provided in Mathematica by the add-on package VectorAnalysis, which needed to be added with the statement << VectorAnalysis'. In subsequent versions, this capacity was made part of the main kernel, though with rather different commands and syntax. The previous package may remain available for a while, but it is officially obsolete and is likely to disappear altogether at some point.
We begin this illustration by defining scalar and vector quantities in each of the three common coordinate systems with the statements

*In*\[1\] := *scar* = x^2 + y^2 + z^2
*Out*\[1\] = x^2 + y^2 + z^2

*In*\[2\] := *scyl* = r^2 + z^2
*Out*\[2\] = r^2 + z^2

*In*\[3\] := *ssph* = r^2
*Out*\[3\] = r^2

*In*\[4\] := *vcar* = \{x, y, z\}
*Out*\[4\] = \{x, y, z\}

*In*\[5\] := *vcyl* = \{r, 0, z\}
*Out*\[5\] = \{r, 0, z\}

*In*\[6\] := *vsph* = \{r, 0, 0\}
*Out*\[6\] = \{r, 0, 0\}

Then, we evaluate the gradients with the statements,

*In*\[7\] := Grad\[scar, \{x, y, z\}\]
*Out*\[7\] = \{2 x, 2 y, 2 z\}

*In*\[8\] := Grad\[scyl, \{r, \[Phi], z\}, "Cylindrical"\]
*Out*\[8\] = \{2 r, 0, 2 z\}

*In*\[9\] := Grad\[ssph, \{r, \[Theta], \[Phi]\}, "Spherical"\]
*Out*\[9\] = \{2 r, 0, 0\}

the Laplacian with the statements

*In*\[10\] := Laplacian\[scar, \{x, y, z\}\]
*Out*\[10\] = 6

*In*\[11\] := Laplacian\[scyl, \{r, \[Phi], z\}, "Cylindrical"\]
*Out*\[11\] = 6

*In*\[12\] := Laplacian\[ssph, \{r, \[Theta], \[Phi]\}, "Spherical"\]
*Out*\[12\] = 6

the divergence with the statements

*In*\[13\] := Div\[vcar, \{x, y, z\}\]
*Out*\[13\] = 3

*In*\[14\] := Div\[vcyl, \{r, \[Phi], z\}, "Cylindrical"\]
*Out*\[14\] = 3

*In*\[15\] := Div\[vsph, \{r, \[Theta], \[Phi]\}, "Spherical"\]
*Out*\[15\] = 3

and, finally, the curl with the statements

*In*\[16\] := Curl\[vcar, \{x, y, z\}\]
*Out*\[16\] = \{0, 0, 0\}

*In*\[17\] := Curl\[vcyl, \{r, \[Phi], z\}, "Cylindrical"\]
*Out*\[17\] = \{0, 0, 0\}

*In*\[18\] := Curl\[vsph, \{r, \[Theta], \[Phi]\}, "Spherical"\]
*Out*\[18\] = \{0, 0, 0\}

*Mathematica* can also evaluate these vector derivatives when the dependence in the first argument on the coordinates is not explicit. For example

*In*\[19\] := Laplacian\[ f[r, \[Phi], z], \{r, \[Phi], z\}, "Cylindrical" \]

*Out*\[19\] = \[f^{(0,0,2)}[r, \[Phi], z] + \frac{f^{(0,2,0)}[r, \[Phi], z]/r + f^{(1,0,0)}[r, \[Phi], z]}{r} + f^{(2,0,0)}[r, \[Phi], z] \]
Here, the superscripts on $f$ indicate how many derivatives with respect to each of the three variables are to be evaluated, e.g.,

$$f^{(2,0,0)}[r, \phi, z] = \frac{\partial^2 f}{\partial r^2} \quad ; \quad f^{(1,1,0)}[r, \phi, z] = \frac{\partial^2 f}{\partial r \partial \phi}$$

where the independence of the mixed derivatives on the order in which the derivatives are evaluated is critical to use of this notation.

We have, of course, mentioned only the commands Grad, Div, Laplacian, and Curl for working with coordinate dependent vectors and scalars. Mathematica also includes a variety of commands to transform the coordinates of a point from one system to another, to calculate the area or volume of appropriately defined regions, to create two and three dimensional graphical displays of vector fields, and many more. You should make it a point to bring up Wolfram Documentation from the NBI HELP menu, search for ‘Vector Analysis’ in that documentation, and study the multitude of available commands. Be aware, also, that Mathematica can apply these commands to functions and vectors in more than three dimensions.

8.9 Packages

To this point, we have supposed that the Mathematica we are using has been configured in the way set up by the default installation, so that all of the functions so far invoked will be available to us once Mathematica has been launched. At some sites, however, Mathematica may have been configured to conserve memory, and some of the features normally loaded automatically may have been left out. At other sites, additional features giving further commands may have been included when Mathematica is started. This flexibility is provided through the use of Mathematica packages, segments of Mathematica code which may or may not be automatically loaded when Mathematica is started up and which define features that will be present only when the package defining those features has been loaded. Any package not loaded automatically can, of course, be explicitly loaded to make its features available. A list of the packages included with the standard Mathematica distribution can be found by accessing the Wolfram Documentation either from the HELP menu in the NBI or on line, searching for ‘Packages’, and opening the link labeled something like ‘Wolfram System Standard Extra Packages’. The resulting page has links to descriptions of the packages. A list of the packages that have been loaded into the current Mathematica configuration is contained in the special variable $\$Packages$, whose value can be displayed by the simple statement\[34\]

\begin{verbatim}
In[9]:= $Packages
Out[9]= ListOfLoadedPackages
\end{verbatim}

If a needed package is not loaded automatically when Mathematica is launched, it can be explicitly loaded with one of the commands\[35\]

\begin{verbatim}
In[23]:= <<PackageName'
In[38]:= Get[ "PackageName" ]
In[79]:= Needs[ "PackageName" ]
\end{verbatim}

Here \texttt{\textasciitilde\textasciitilde} and \texttt{Get} are synonymous and will load the package directly, even if it has previously been loaded. In contrast, \texttt{Needs} will load the package only if it has not already been loaded.

Beyond the standard packages, individual users can create additional packages for special purposes. Thus, if we are lucky, we might be able to find a user-contributed package with special features

\[34\] The list output by this statement may vary from site to site, depending on how Mathematica is configured. It will also reflect the inclusion of any packages explicitly loaded during the session.

\[35\] Note that the tic following \texttt{PackageName} in the first statement is a forward tic.
of value to a particular exploration. In the absence of such luck, we could also create our own package to serve those purposes. The details of writing such packages are laid out in the Mathematica manuals but are beyond the scope of this book.

With each new version of Mathematica, some features that had been provided with add-on packages may be moved to the main kernel, in which case those features will be available without explicitly loading the package. Indeed, prior to Version 9, the capacity for evaluating several vector derivatives described in Section 8.8.10 was available only if a package named VectorAnalysis was loaded explicitly. Starting with Version 9, those features were built into the main kernel, though the names of the commands and their syntax suffered revision. For backward compatibility, the original package is retained in subsequent versions, at least for a while, but it may in time be completely removed. Further, loading the package will also generate a message alerting the user that the package has become obsolete.

### 8.10 Loops, Logical Expressions, and Conditionals

Among the most ubiquitous programming structures is the loop, which provides a means by which a statement or block of statements can be executed some number of times, typically with small changes controlled by a loop index. Several such structures are available in Mathematica. The simplest is the For loop, which we have already met in Sections 8.8.2 and 8.8.5. Briefly, the statements

```plaintext
In[1]:= x = {0, 0, 0, 0, 0, 0, 0, 0, 0, 0 };
Out[1]= {0, 0, 0, 0, 0, 0, 0, 0, 0, 0 }
In[2]:= For[ i = 1, i <= 10, i = i + 2, x[[i]] = i^2 ];
In[3]:= x
Out[3]= {0, 0, 9, 0, 25, 0, 49, 0, 81, 0 }
```

create a 10-element list of zeros and then assign the squares of the odd integers to the odd elements of that list. Note that the C-like structure `i++` can be used instead of the expression `i = i + 1` if the index is to be incremented by 1.

In essence, the For loop requires a priori knowledge of how many times the loop will be executed. For cases when that knowledge is not in hand, Mathematica makes available the While loop, whose general syntax has the form

```plaintext
While[ (condition), (block of statements) ]
```

In this loop, the block of instructions must assume responsibility for changing the condition to a value that will stop the loop. More specifically, these loops are illustrated in the statements

```plaintext
In[4]:= x = {0, 0, 0, 0, 0, 0, 0, 0, 0, 0 };
In[5]:= i = 1;
In[6]:= While[ i < 11, x[[i]] = i^3; i = i+2 ];
In[7]:= x
Out[7]= {1, 0, 27, 0, 125, 0, 343, 0, 729, 0 }
```

Note that,

- The block of statements in this structure starts after the first comma and continues until the closing bracket.
- There is no punctuation after the last statement in the block of statements.
- Individual statements in that block are separated by semicolons.
The statement incrementing $i$ is essential so that the initially true condition $i < 11$ is ultimately toggled to false and the loop terminates.

Finally, a word about composite conditions: The illustrations above made use of (single) conditions like $i < 11$, which will evaluate to true or false depending on the value of $i$. Indeed, the Mathematica function \texttt{Boole}, for example,

\begin{verbatim}
In[8]:= i := 5:
In[9]:= { Boole[ i > 10 ], Boole[ i < 10 ] }
Out[9]= [0, 1]
\end{verbatim}

allows the testing of a logical condition. In this Boolean context, 0 equates to false and 1 to true. The six logical operators $==, <, >, <=, >=, $ and $!$ can be used for test for equal, less than, greater than, less than or equal, greater than, and unequal. In some situations, one may need to (1) combine two or more conditions into a composite condition or (2) negate a condition. \texttt{Mathematica} makes available the logical operators $\&\&$, $|\|$, and $！$, and the logical functions \texttt{Xor} (exclusive or), \texttt{Nand}, and \texttt{Nor} to facilitate combining conditions and negating conditions. More specifically, with $\langle C_1 \rangle$ and $\langle C_2 \rangle$ individual conditions

- $(\langle C_1 \rangle \&\& \langle C_2 \rangle)$ will be true if $\langle C_1 \rangle$ and $\langle C_2 \rangle$ are both true and false otherwise,

- $(\langle C_1 \rangle |\| \langle C_2 \rangle)$ will be true if either $\langle C_1 \rangle$ or $\langle C_2 \rangle$ is true or both $\langle C_1 \rangle$ and $\langle C_2 \rangle$ are true and false otherwise, and

- $\texttt{Xor}[\langle C_1 \rangle, \langle C_2 \rangle]$ will be true if either $\langle C_1 \rangle$ or $\langle C_2 \rangle$ is true and false if neither condition is or both are true, and

- $!\langle C_1 \rangle$ will be true if $\langle C_1 \rangle$ is false and false if $\langle C_1 \rangle$ is true.

For example, the coding

\begin{verbatim}
In(10):= x = { 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 };
In(11):= i = 1;
In[12]:= While[ (i > 0 && i < 11), x[[i]] = i^3; i = i+1 ]
In[13]:= x
Out[13]= [1, 8, 27, 64, 125, 216, 343, 512, 729, 1000]
\end{verbatim}

illustrates a composite condition controlling the \texttt{While} statement. Note that parentheses around the composite condition can in this case be omitted, but they—and perhaps parentheses around some internal pieces—may be necessary in some cases.

Logical conditions appear not only to control loops but also to structure branches in a sequence of statements. As with most programming languages, \texttt{Mathematica} also provides the function \texttt{If} for this purpose. Its fundamental syntax is

\begin{verbatim}
\texttt{If[ \langle condition \rangle, \langle statements if true \rangle, \langle statements if false \rangle ]}
\end{verbatim}

though the second block of statements can be omitted if nothing is to be done when the condition is false. Thus, for example, the statement

\begin{verbatim}
For[ i=1, i<=10, i++, If[ x[[i]] < 0, x[[i]] = -x[[i]] ] ]
\end{verbatim}

will replace each negative element in a list $x$ with the corresponding positive value, the statement

\begin{verbatim}
For[i = 1, i<=10, i++, If[ x[[i]]> 10, x[[i]] = 10 ] ]
\end{verbatim}
will replace all values greater than 10.0 with the value 10.0, and the statement

\[
\text{If[ } a \geq 0, b = a, b = -a \text{ ]}
\]

will set \( b \) equal to the absolute value of \( a \) (though the function \texttt{Abs} will do so more easily). Multiple statements in either the true clause or the false clause will be separated by commas and enclosed within braces, as in

\[
\text{If[ } \langle \text{condition} \rangle, \{ \langle \text{statement} \rangle, \langle \text{statement} \rangle, \ldots \} \\
\{ \langle \text{statement} \rangle, \langle \text{statement} \rangle, \ldots \} \text{ ]}
\]

For selection among more than two options, the statement would have to be nested with care. All conditions may be single or composite.

As is the case in most languages, loops and conditional structures can be nested, though constructing the syntax correctly can sometimes be challenging.

### 8.11 Command Files

In addition to responding to statements supplied interactively, Mathematica can execute statements read from a command file. To begin, we create a text file containing the desired statements just as they would be typed interactively. If that file is then stored in the default directory\(^{36}\) or somewhere in Mathematica’s search path,\(^ {37}\) execution of the command(s) contained within it would then be specified by typing either of the statements

\[
\texttt{\textless \textless FileName} \quad \text{or} \quad \texttt{Get["FileName"]}
\]

when Mathematica is ready to accept input. There is no default file type.\(^ {38}\) As the command file is read and executed, the statements themselves will \textit{not} be displayed on the screen. Whether the statements are terminated with semicolons or not, the output will also not be displayed on the screen (except for the last statement in the file, whose output will be displayed if it ends with no semicolon and suppressed if it ends with a semicolon). Further, no line labels will be attached to any of the intermediate results, and the output from the last statement in the file (if any) will be given the label Out\([n]\), where \( n \) is the identifying number of the statement requesting the reading of the file. Only if intermediate results are explicitly printed (with the command \texttt{Print[]} ) will these results be displayed; only if intermediate results are explicitly bound to a variable will those results be available once execution of the command file has been completed. Except for the absence of the In and Out labels, the Mathematica environment after the command file has been executed is identical to the environment that would have existed had the statements been entered interactively.

Sometimes, creation of a command file facilitates repeated use of quantities not defined in standard Mathematica. Use of such a file can also facilitate the debugging of a sequence of statements, since—when an error is reported—we need only edit the file appropriately and resubmit it to Mathematica; repeated typing of the correct parts of the sequence is not then necessary. To illustrate matrix multiplication, for example, we could create the file \texttt{testmultiply.m} containing the statements listed in Table 8.1.\(^ {39}\)

Then, assuming that this file has been stored in the default directory,\(^ {40}\) we would execute the statements it contains by submitting to Mathematica one or the other of the statements

---

\(^{36}\)See Section 8.18.3.

\(^{37}\)See Section 8.18.4.

\(^{38}\)Conventionally, text files containing Mathematica statements are given names with file type .m. Further, the types .nb and .mx are conventionally used for files containing Mathematica notebooks and Mathematica definitions in a special format, respectively.

\(^{39}\)Comments in Mathematica code are introduced with the symbol (* and terminated with the symbol *).

\(^{40}\)See again Section 8.18.3.
Table 8.1: The Mathematica command file testmultiply.m.

(* Command file testmultiply.m *)

(* Enter and display matrix *)
mat1 = { {1,2}, {3,4} };
Print[ MatrixForm[ mat1 ] ]
(* Evaluate and display transpose *)
mat2 = Transpose[ mat1 ];
Print[ MatrixForm[ mat2 ] ]
(* Evaluate and display matrix product *)
Print[ MatrixForm[ mat1.mat2 ] ]
(* Evaluate and display element-by-element product *)
Print[ MatrixForm[ mat1*mat2 ] ]

Table 8.2: The Mathematica command file crossdot.m.

(* crossdot.m *)

(* Command file to define functions to evaluate the cross and dot products of two three-component vectors. *)
lucross[a_,b_] :=
{ a[[2]]*b[[3]] - a[[3]]*b[[2]],
a[[3]]*b[[1]] - a[[1]]*b[[3]],
a[[1]]*b[[2]] - a[[2]]*b[[1]] };
ludot[a_,b_] := a[[1]]*b[[1]] + a[[2]]*b[[2]] + a[[3]]*b[[3]];

<< testmultiply.m or Get[ "testmultiply.m" ]

All output produced by the Print statements will be displayed on the screen but only the expressions explicitly bound to mat1 and mat2 will be available by those names once execution of the file is completed.

As a second example, suppose that we frequently had two three-component lists defining two three-dimensional vectors and that we wanted to be able easily to evaluate their cross and dot products. We might create a command file listed in Table 8.2. Here, beyond the comments, the first four lines—actually a single statement—define the function lucross, which takes two three-component lists as arguments and returns a three-component list whose components are those of the cross product of the two vectors. The last line defines the function ludot, which takes the same arguments but returns a scalar value equal to the dot product of those two lists treated as three-dimensional vectors. This file is named crossdot.m and can be accessed in the directory $HEAD/mathematica. Thus, for example, the statement

[^41]: See the note at the beginning of this chapter.
will load this file from the directory associated with this book, defining the two functions ludot and lucross for subsequent use. \footnote{Be aware, as well, that the dot product of two vectors can also be evaluated with the non-commutative infix operator \texttt{\_\_\_\_\_\_}, which will actually provide the sum of the products of corresponding components regardless of the dimensionality of the vectors. Further, in addition to the several vector derivatives already introduced in Section 8.8.10, \texttt{\textit{Mathematica}} includes the commands \texttt{Dot} and \texttt{Cross}, which are more carefully structured than \texttt{ludot} and \texttt{lucross} to evaluate dot and cross products.}

Finally in this section, we note that \texttt{Mathematica} also provides means by which we can display the contents of a command file conveniently and easily from within \texttt{Mathematica}. To display the file \texttt{crossdot.m}, we submit to \texttt{Mathematica} the statement

\begin{verbatim}
FilePrint[ "crossdot.m" ] or FilePrint[ "$HEAD/mathematica/crossdot.m" ]
\end{verbatim}

Here, the first form applies if the file is in a directory in the search path and the second form refers to the file in the directory associated with this book. The contents of the file will be displayed in the NBI or, if you are using the TBI, in the \textit{Shell} window from which \texttt{Mathematica} was launched.

### 8.12 High-Resolution Graphing

We illustrate the production of high resolution graphics by plotting the solution for the damped harmonic oscillator obtained in Section 8.8.9. First, we reclaim the solution with the statement

\begin{verbatim}
In[1]:= soln3 = x0 \ast \text{Exp}\[-\\Gamma \ast t\] \ast (\Gamma \text{Sin}[a \ast t]/a + \text{Cos}[a \ast t])
\end{verbatim}

\begin{verbatim}
Out[1] = e^{-t\Gamma}x_0\left(\text{Cos}[a t] + \frac{\Gamma \text{Sin}[a t]}{a}\right)
\end{verbatim}

In addition, we calculate and display the velocity with the statement

\begin{verbatim}
In[2]:= soln3dot = Simplify[ D[ soln3, t ] ]
\end{verbatim}

\begin{verbatim}
Out[2] = -e^{-t\Gamma}x_0(a^2 + \Gamma^2)\text{Sin}[a t]
\end{verbatim}

Then, recognizing that $\gamma$ and $a$ are not independent ($a^2 = k/m - \gamma^2$) and hence that assigning values to $\gamma$ and $a$ in effect fixes $k/m$, we use the statements

\begin{verbatim}
In[3] := \text{\Gamma} = 0.1; a = 1.0; x0 = 1.0;
\end{verbatim}

to assign specific values to $\gamma$, $a$, and $x_0$. Finally, we create an inline graph showing position and velocity with the statement

\begin{verbatim}
In[4]:= Plot[ {soln3,soln3dot}, {t,0.0,20.0} ]
\end{verbatim}

which plots the identified expressions over the intervals $0.0 \leq t \leq 20.0$ in $t$ and $-1.0 \leq y \leq 1.0$ in $y$ (the vertical coordinate). The graph is produced as a coarse character plot if we are working in the \texttt{Mathematica} TBI and as a smooth plot in the NBI. \footnote{The number of points distributed over the range of the independent variable is determined in \texttt{Mathematica} to keep the amount of turning from segment to segment small. Numerous options—see the \texttt{Mathematica} manuals—give the user the ability to override the defaults.}

Alternatively, we could modify the resulting graph by exploiting a few of the many options supported by the command \texttt{Plot}. The statement

```mathematica
\begin{verbatim}
Get[ "$HEAD/mathematica/crossdot.m" ];
\end{verbatim}
```
8.12. HIGH-RESOLUTION GRAPHING

Figure 8.1: Graph of position and velocity for a damped oscillator with $\gamma = 0.1$, $a = 1.0$, $x(0) = x_0 = 1.0$, and $v(0) = 0.0$. Here, the solid and dashed lines show $x(t)$ and $v(t)$, respectively.

\begin{center}
\textbf{Damped Oscillator}
\end{center}

\begin{center}
\includegraphics[width=\textwidth]{damped_oscillator_graph.png}
\end{center}

\texttt{In[5]:=} \texttt{Plot[ \{soln3, soln3dot\}, \{t,0.0,20.0\}, PlotRange->\{-1.0,1.0\},
PlotPoints->100, AxesLabel->\{"t","x,v"\},
PlotLabel->StyleForm[ "Damped Oscillator",
FontSize->16, FontFamily->"Times", FontWeight->"Bold" ],
PlotStyle->{ \{Thickness[0.01],Black\},
\{Thickness[0.01],Black,Dashing[{0.02,0.02}]\} },
Epilog->{Text["\[Gamma]=0.1, a=0.1", \{15.0, 0.75\}] ]}

\texttt{In[5]:=} \texttt{Plot[ \{soln3, soln3dot\}, \{t,0.0,20.0\}, PlotRange->\{-1.0,1.0\},
PlotPoints->100, AxesLabel->\{"t","x,v"\},
PlotLabel->StyleForm[ "Damped Oscillator",
FontSize->16, FontFamily->"Times", FontWeight->"Bold" ],
PlotStyle->{ \{Thickness[0.01],Black\},
\{Thickness[0.01],Black,Dashing[{0.02,0.02}]\} },
Epilog->{Text["\[Gamma]=0.1, a=0.1", \{15.0, 0.75\}] ]}

for example, will produce the same graph but with

\begin{itemize}
\item the vertical axis running from $-1.0$ to $+1.0$,
\item at least 100 points spaced over the range of the independent variable (the default is 25),
\item the indicated labels on the horizontal and vertical axes,
\item the indicated title in 16-point, bold type, Times font on the graph as a whole,
\item the first expression (but not the axes) displayed black with a line of thickness heavier than the default (i.e., with a thickness set to the indicated fraction of the height of the graph as a whole) and the default (solid) linestyle and the second expression displayed black with the same line thickness but with a dashed line style with alternate dashes and gaps, each of size 0.02 units, and
\item the text $\gamma = 0.1, a = 0.1$ displayed on the face of the graph and centered at the point [15.0, 0.75].\footnote{See the \textit{Mathematica} manuals at Epilog and Text for details on the way to specify text displayed on a graph.}
\end{itemize}

These (and other) options to the command \texttt{Plot} are fully described in the \textit{Mathematica} manuals. The graph resulting from the statement on line \texttt{In[5]} is shown in Fig. 8.1.

We could explore other features of this motion by, for example, changing the value of $\gamma$ and reploting the graph with the statements

44See the \textit{Mathematica} manuals at Epilog and Text for details on the way to specify text displayed on a graph.
Figure 8.2: Graph of velocity versus position for a damped oscillator with $\gamma = 0.2$, $a = 1.0$, $x(0) = 1.0$, and $v(0) = 0.0$.

In[6]:= $\text{\textbackslash Gamma} = 0.2$
In[7]:= Plot[ {soln3, soln3dot}, {t,0.0,20.0}, PlotRange->{-1.0,1.0},
PlotPoints->100, AxesLabel->"t","x,v"],
PlotLabel->"Damped Oscillator",
FontSize->16, FontFamily->"Times", FontWeight->"Bold",
PlotStyle->{ Thickness[0.01],Black,
 {Thickness[0.01],Black,Dashing[{0.02,0.02}]} }
]

or, for these new values of the parameters, we could produce a graph of the trajectory in the phase plane (soln3dot versus soln3) with the statements

In[8]:= ParametricPlot[ {soln3, soln3dot}, {t,0.0,20.0},
PlotRange->{ -1.0, 1.0 },
PlotStyle->{ Thickness[0.01],Black},
AspectRatio->1.0 ]

Here, the new option AspectRatio permits us to tell *Mathematica* to use a plotting area that is square (aspect ratio 1.0). The resulting graph is shown in Fig. 8.2.

*Mathematica* can, of course, also produce surface plots and contour plots in several different ways. The simplest and quickest such plots are produced with statements like

In[9]:= mode23 = Sin[2*Pi*x] * Sin[3*Pi*y]
In[10]:= Plot3D[ mode23, {x,0.0,1.0}, {y,0.0,1.0},
PlotPoints->25, ColorFunction->GrayLevel ]
In[11]:= lvls = Range[ -9,9,1]/10.0;
In[12]:= ContourPlot[ mode23, {x, 0.0, 1.0}, {y, 0.0, 1.0},
AspectRatio->1.0, ContourShading->False,
Contours->lvls, PlotPoints->50 ]
In[13]:= Quit[]
Figure 8.3: Surface plot of the function $z(x, y) = \sin(2\pi x) \sin(3\pi y)$.

Line In[9] here defines the expression to be displayed. Line In[10] produces a surface plot with 25 points (default 15) along each axis and coloring set to a gray scale (since we anticipate black and white printing in this book). Finally, line In[11] defines the values at which contour lines are too be drawn (from $-0.9$ to $0.9$ in steps of $0.1$) and line In[12] produces a contour map at those levels (option Contours) with shading turned off, the aspect ratio adjusted so that the display will be square, and the function evaluated at 50 points along each axis. The resulting graphs are shown in Figs. 8.3 and 8.4.

We have, of course, only mentioned a few (Plot, ParametricPlot, Plot3D, and ContourPlot) of the functions and only some (PlotRange, PlotPoints, AxesLabel, PlotLabel, PlotStyle, FontSize, FontWeight, ColorFunction, AspectRatio, ContourShading, and Contours) of the options that modify the behavior of these functions. Several other functions for producing graphical displays and numerous additional options are available. You should make it a point to read about these functions and the associated options in the Mathematica manuals. In particular, note that the statement `Options[Function]`\textsuperscript{45} will return a list of all of the options that affect Function, attaching to each its current default value.

8.13 Making Hard Copy . . .

The method used to make hard copy of all or part of a Mathematica dialog depends on whether the TBI or the NBI is in use; whether the material to be output is textual or graphical; whether the graphical material is inline or in a separate window; whether the output is to go directly to a printer or to a PostScript, ASCII, HTML, or . . . file; and whether we are working on a UNIX, Windows, or Macintosh platform. In the following subsections we address each of several possibilities in turn.

\textsuperscript{45}See Section 8.18.5 for a fuller discussion.
8.13.1 ... from the NBI

The menu structure and procedures to print text and graphics from the NBI may differ from version to version and platform to platform. For obtaining hard copy in the NBI, several items in the FILE menu are relevant:

- The option ‘Print...’ opens a pop-up window in which to specify a printer and, if the OPTIONS or PREFERENCES menu is opened, select various parameters (page range to be printed, number of copies to be printed, treatment of color in the output, ...). In Windows, a capacity to print a selection from the active notebook is provided as an option in the ‘Page Range’ panel of the ‘Print...’ option in the FILE menu; in UNIX, a capacity to print a selection from the active notebook is provided in the separate option ‘Print Selection...’ in the FILE menu. In either case, the option will be available only if a portion of the notebook has, in fact, been selected.

- The option ‘Printing Settings→’ from the FILE menu will open a menu offering several options, including:
  - ‘Page Setup’, which opens a pop-up window in which to specify paper size, orientation (portrait or landscape), and margins.
  - ‘Printing Options’, which opens a pop-up window in which to specify whether to print cell brackets, registration marks, and a few other features.
  - ‘Headers and Footers’, which opens a pop-up window in which to specify the content and position of desired page headers and footers.
  - ‘Show Page Breaks’, a check box, to select whether page breaks should be displayed or hidden in the NBI.
  - ‘Printing Environment’, which opens a pop-up window in which to select among the options ‘Working’, ‘Slide Show’, or ‘Printout’.

---

46This description has been worked out in LINUX and Windows environments. The same features will be available in other environments, but the details of the behavior may be slightly different. Further, the details here described apply specifically to Mathematica 11. Other versions may make these capabilities available on other ways.

47‘Print to file’ is among the options but offers fewer formats than ‘Save’, ‘Save As’, and ‘Save Selection As’.

Figure 8.4: Contour map of the function $z(x, y) = \sin(2\pi x)\sin(3\pi y)$. 
8.13. MAKING HARD COPY...

- The option ‘Print Preview...’ displays on screen what will ultimately be printed. That on-screen display will be produced in an Adobe Reader window from which a PDF file can, of course, be saved.

- The options ‘Save...’, ‘Save As...’, and ‘Save Selection As...’, each of which brings up a browser in which you can identify the directory in which all or the selected part of a notebook should be saved, the name to give the resulting file, and the file type to be used. Available file types include tex, ps, eps, pdf, and html, though not all types are available with all options. In particular, the filetype ps is available with ‘Save...’ and ‘Save As...’ and the file type eps is available with ‘Save Selection As’. The resulting file can then be printed directly or incorporated in a more extensive document.

These routes to output information include the Mathematica-generated labels for all included elements of the notebook.

Mathematica also provides a facility for writing a description of a specified graphic object (not the entire notebook) into a file. We simply need to inform Mathematica of the form of the driver it is to use in creating the file, the name it is to give the file, and any controlling options. For example, with graph the label of the graphics display to be output to a file, either of the statements

\[
\text{Export[ "trial.eps", graph, "EPS"]}
\]

which determines the format of the output file from the third argument, or

\[
\text{Export[ "trial.eps", graph]}
\]

which determines the format of the output file from the extension in the file name,\(^48\) will tell Mathematica to convert the graphic object identified by the variable graph into encapsulated PostScript and write that description to the file trial.eps in the default directory.\(^49\) When displayed on a computer screen with ghostview, an EPS file creates a display that includes a number of controlling buttons surrounding the graph itself. These surrounding buttons are not included in PDF or JPEG files, nor do they appear, for example, when the EPS file is incorporated in a \LaTeX{} document. This route to exporting a graph does not include Mathematica-generated output labels in the file.

Note, however, that in some contexts Mathematica may not always transcribe all characters in the text in the on-screen graphic (e.g., in labels on the axes or in titles) into the proper PostScript characters. Unless some Mathematica-specific fonts have been made available to the program that will interpret the PostScript file, opening and closing parentheses, the forward slash, and probably other characters may be rendered in unexpected forms when the resulting PostScript file is printed or viewed on the screen. In exporting graphics, we must either be sure that the appropriate Mathematica fonts are known by the program(s) that will display the file, or we must constrain Mathematica as it produces the graphic by limiting its choice of fonts for use in the graphic. The former route is complicated (and necessary if, for example, Greek letters are to be included in the exported text).\(^50\) While not always satisfactory, the latter route is far simpler; it involves the stipulation of an additional option in the command producing the graphic. For the sake of a quick illustration, we simply note that the statements

\[
\text{graph = Plot[ Sin[t], \{t, 0, 10\}, AxesLabel->\{"sin(t)\", \"x/a\"\},}
\text{ FormatType->OutputForm \} }
\text{Export[ "trial.eps", graph, "EPS"]}
\]

will produce an encapsulated PostScript file in which the parentheses and the slash in the labels on the axes will be properly rendered, even when the specific Mathematica fonts are not available to the program that will display trial.eps.

\(^{48}\)Other available formats include "PDF" (file type pdf and "JPEG" (file type jpeg. A full list is stored in the system variable $\text{ExportFormats}$.\(^ {49}\)See Section 8.18.3.\(^ {50}\)For details, see the Mathematica manuals.
8.13.2 ... from the TBI

Printing of text displayed in the TBI is most simply accomplished by using the cutting and pasting capabilities of the operating system to transfer that character information into a text file, which can then be edited in a standard text editor and sent to an available printer with the standard commands to the operating system (or perhaps by selecting appropriate options from a menu available with the editor).

Graphics commands like \texttt{Plot} may not produce very satisfactory on-screen results when executed in the TBI. Even so, if it is given a variable name, a graphic object can still be written into a file as high-resolution graphic for subsequent printing. As in the previous section, we simply need to inform \textit{Mathematica} of the form of the driver it is to use in creating the file, the name it is to give the file, and any controlling options. For example, either of the statements

\begin{verbatim}
Export[ "trial.eps", graph, "EPS" ]
\end{verbatim}

or

\begin{verbatim}
Export[ "trial.eps", graph ]
\end{verbatim}

will tell \textit{Mathematica} to convert the graphic object identified by the variable \texttt{graph} into encapsulated PostScript and write that description to the file \texttt{trial.eps} in the default directory.

8.14 Output in \LaTeX{} Format

A particularly useful feature of \textit{Mathematica} is its ability to write \LaTeX{} versions of expressions to the screen or into a file. Full details are laid out in the \textit{Mathematica} manuals. In brief, the statement\footnote{Note the \texttt{upper-case X} in the third character.}

\begin{verbatim}
TeXForm[ expr1 ]
\end{verbatim}

for example, converts the specified expression into \LaTeX{} format and displays it on the screen (from where standard cut and paste operations—when they are available—can move the description to wherever we might want it). Alternatively, either of the statements

\begin{verbatim}
TeXForm[ expr1 ] >> FileName
Put[ TeXForm[ expr1 ], "FileName" ]
\end{verbatim}

will create a new (ASCII) file and write the \LaTeX{} description of the indicated expression into that file, \textit{overwriting} any existing file by the same name. To \textit{append} the \LaTeX{} description of the expression to an \textit{existing} file, one or the other of the statements

\begin{verbatim}
TeXForm[ expr1 ] >>> FileName
PutAppend[ TeXForm[ expr1 ], "FileName" ]
\end{verbatim}

is necessary. Once created, this file can, of course, be edited or incorporated into other \LaTeX{} documents.

Automatic translation of arbitrary expressions into \LaTeX{} is a daunting task. While fairly sophisticated, the \textit{Mathematica} translator is not perfect. Edits will often be necessary to “correct” the translator’s glitches. In particular, the translator writes only the \LaTeX{} description of the
8.15. ANIMATION

equation to the screen or into the file. It does not place the equation in a \LaTeX equation environment or otherwise embellish the description with additional \LaTeX constructs that may be necessary for creation of the ultimate document. These formatting instructions will have to be supplied through manual editing of the file created by Mathematica. Further, even more subtle edits may be necessary.\footnote{For example, output of the \textit{Mathematica} expression $\sin (x+y)/\cos (x-y)$ with \texttt{TeXForm} yielded the correct but unintended result $\sin (x+y) \sec (x-y)$. Note also the unnecessary spaces after \texttt{sin} and \texttt{cos}.} Conveniently, variables named $\\texttt{\textbackslash Gamma}$ and $\\texttt{\textbackslash CapitalGamma}$, for example, will be properly translated into $\gamma$ and $\Gamma$.

Even more generally, if we have created a \textit{Mathematica} notebook in the NBI, we can output that notebook in its entirety (including the textual sections) into a \LaTeX file by

- Selecting ‘Save as …’ from the \texttt{FILE} menu,
- Selecting the \LaTeX option in the text entry box labeled ‘Files of type’,
- Specifying the desired file name in the text entry box labeled ‘File name’,
- Browsing to the desired directory, and
- Clicking ML on ‘Save’.

The file will be written into the specified directory. The resulting file is a complete \LaTeX source file with all necessary packages and new command definitions in its preamble. If the content of this file is to be incorporated into another \LaTeX source file, some of the commands at its beginning and end will have to be stripped off but any \texttt{\usepackage} and \texttt{\newcommand} commands not already in the destination file must be copied from the \textit{Mathematica}-produced file.\footnote{\LaTeX files written by some earlier versions of \textit{Mathematica} make use of several \textit{Mathematica}-specific \LaTeX commands which are defined in the \textit{Mathematica}-supplied \LaTeX package named \texttt{notebook2e} and the \LaTeX-supplied package named \texttt{latexsym}. The files defining these packages must be accessible on your computer system before a \textit{Mathematica}-produced \LaTeX source file can be processed and printed. The file \texttt{latexsym.sty} is probably available without effort on your part. If the file \texttt{notebook2e.sty} has not been placed where \LaTeX can find it automatically, you may have to copy it from the directory \texttt{\$MATHEMHEAD/SystemFiles/IncludeFiles/Tex/texmf/tex/latex/wolfram}, in which it was placed when \textit{Mathematica} was installed on your system.}

8.15 Animation

The \textit{Mathematica} command \texttt{Animate} facilitates the creation of animated displays in the NBI. The simple statement

\begin{verbatim}
Animate[ Plot[Sin[Pi*x]*Cos[2*Pi*t], {x,0.0,1.0},
        PlotRange->{ {0.0,1.0},{-1.0,1.0} }], {t, 0.0, 1.0, 0.05} ]
\end{verbatim}

will, for example, display the fundamental mode of a vibrating string, keeping the axes the same size throughout all frames. A more complicated animation, display of the 23 mode of a square membrane, is produced by the statement

\begin{verbatim}
Animate[ Plot3D[Sin[2*Pi*x]*Sin[3*Pi*y]*Cos[2*Pi*t], {x,0.0,1.0}, {y,0.0,1.0},
        PlotRange->{ {0.0,1.0},{0.0,1.0},{-1.0,1.0} }], {t,0.0,1.0,0.05},
        AnimationRate-> 0.1]
\end{verbatim}

In both cases, the button labeled ‘Pause’ in the window displaying the animation will stop the animation and allow subsequent statements to be submitted to \textit{Mathematica}.

Unfortunately, the command \texttt{Animate} has too many possibilities and options for full discussion here. Details can be found in the \textit{Mathematica} manuals.
8.16 Using the Notebook

Within the NBI, the window in which Mathematica commands are entered has full capabilities as a notebook, which means in particular that it can contain textual sections, e.g., documentation and motivation, as well as executable Mathematica commands. Notebooks are constructed out of cells, which may contain executable Mathematica statements (and the associated output)—the only cells we have so far seen—or may contain text. For purposes of organizing larger documents, these cells can themselves be assembled into sections and subsections, much in the way a conventional book would be assembled.

Suppose we already have on screen a Mathematica notebook containing a few Mathematica statements and their output. Note, in particular, that each statement and its output embraces two cells, one containing the input statement and the other containing the associated Mathematica response. The input cell is marked at the right side of the Notebook window with a closing square bracket whose upper corner contains also a diagonal line, and the output cell is similarly marked except that the bracket also has a short horizontal line below the diagonal line. Further, the two cells are themselves grouped together at the right side of the Notebook window with a closing square bracket that has no adornments at its corners. Finally, note that double clicking ML on the outer square bracket will “close” the pair, hiding the output and leaving only the input statement. In this arrangement, the outer bracket now has half an arrowhead at its lower end to indicate that the cell has been closed. Double clicking ML on the outer bracket when it is in this state will reopen the cell.

Now, suppose that we want to precede each statement with an explanatory paragraph. We could achieve that end for one of the statements by

1. Moving the cursor to the position at which a new cell is desired, placing it so that it becomes a horizontal I-beam.
2. Clicking ML to create a horizontal line (cell insertion bar) across the screen.
4. Typing the desired text, using the various options in the Format menu to specify font, type size, type style, alignment, .... As you start typing, a new cell will be created for the typed text. This cell will be bracketed on the right with a square bracket that has a short horizontal line below its top bar—the identifier of a text cell.

This process can, of course, be repeated to provide a textual description for each of the remaining statements.

In essence, the previous paragraph provides all we need to know to add documentation as an intrinsic part of a notebook. As our notebooks grow in size, however, we may wish to organize their contents into appropriate structural units and provide titles. To those ends, we need simply notice that the Style submenu of the Format menu contains numerous items beyond the one introduced in the previous paragraph. Positioning the cursor at the point of a desired textual insertion, clicking ML to create a horizontal line (cell insertion bar) across the screen, and then selecting some other item (Title, Subtitle, Section, Subsection, ...) from the Style submenu of the Format menu will then set Mathematica to create a cell of the selected style when we begin typing. Further, Mathematica will select a type size and weight appropriate to the selected cell style and, for sections and subsections, will place a marker at the beginning of the line to draw attention to its status in clarifying the structure of the document. Note also that, as a document is divided into sections and subsections, the bars along the right edge become more numerous, reflecting the divisions created by the inserted sections and subsections.

For example, to add a title to the entire document, we create an insertion bar at the very beginning of the document, select ‘Title’ from the Style submenu of the Format menu, and type
the text of the title. Then, selecting this cell by highlighting its bracket at the right edge of the screen, we can adjust the font, type style, type size, alignment, and several other properties by appropriate selections from submenus accessible in the Format menu.

Mathematica output in this notebook can, of course, always be quickly regenerated by executing the statements. Frequently, we may wish to suppress the presence of that output in the actual notebook. We can

- remove selected items of output by selecting the items to be removed and then selecting ‘Cut’ from the Edit menu.
- remove all output from the notebook by selecting ‘Delete All Output’ from the Cell menu.

Finally, we note that, after all the work of creating a notebook, the notebook itself can be saved by selecting ‘Save’ from the File menu and, in the resulting pop-up window, specifying the file name (conventional file type .nb) and the directory, and clicking ML on the ‘Save’ button. In a later session, the worksheet can be loaded into Mathematica by selecting ‘Open’ from the File menu and identifying the file to be opened in the pop-up browser.

This short section can, of course, not do full justice to the wide spectrum of features of Mathematica’s notebook interface. You must work with it, explore the menus in the NBI, and read the Mathematica manuals

### 8.17 Customizations

When Mathematica is started up, the configuration both of the kernel and of the front-end is defined by a number of default specifications built into the program, by statements in an assortment of system-wide initialization files and by statements in several user-created initialization files. Specifically,\(^{54}\)

- The file $BaseDirectory/Kernel/init.m configures the Mathematica kernel for all users and the file $BaseDirectory/FrontEnd/init.m configures the Mathematica front-end for all users. These files are maintained by the system manager and can be viewed but not edited by individual users.
- The file $UserBaseDirectory/Kernel/init.m modifies the configuration of the kernel established by the system-wide file and must be explicitly created by the individual user.
- The file $UserBaseDirectory/FrontEnd/init.m modifies the configuration of the front-end established by the system-wide file but is not created directly by the user. Instead, Mathematica maintains this file to reflect changes made via statements executed by Mathematica.

These files are simply command files—see Section 8.11—that can contain any statement that might otherwise be typed interactively. Be aware, however, that none of the statements or their responses will be displayed as the initialization file(s) is (are) read. Even so, for example, variables to which values are assigned and the features defined by packages read in will be available when reading of the file has been completed. As Mathematica is started, the system-wide files are executed and then the user-created file is executed. Thus, the user-created file can be used to override statements in the system-wide file, should that be desired. Together, these files—if they exist at all—define the default configuration set up by Mathematica when finally it is ready to accept input.

---

\(^{54}\)Here, the system variable $BaseDirectory identifies the top directory in the hierarchy in which the Mathematica structure is stored and the system variable $UserBaseDirectory identifies the top directory in which user-created configuration files are stored. Both of these directories may be hidden.
8.18 Miscellaneous Occasionally Useful Tidbits

In this section, we comment on a number of other features that are occasionally useful. Full details, of course, are available in the Mathematica manuals.

8.18.1 Turning off Messages

Many of Mathematica's messages can be suppressed by executing the command \texttt{Off[...] } and turned back on by executing the command \texttt{On[...]}. Here, the ellipses stand for the name of the message. For example, to turn off display of the message that alerts you that a symbol you have just typed is \textit{nearly} the same as some other symbol that you may have intended to type, you can execute the statement

\texttt{Off[General::Spell1]}

though full turning off of these messages may also require execution of the statement

\texttt{Off[General::Spell]}

These messages can be turned on again with the statements

\texttt{On[General::Spell1]}
\texttt{On[General::Spell]}

The system variable $\texttt{MessageGroups}$ stores a list of the existing message groups (the piece of the argument before the double colon).

8.18.2 Specifying Directories

Different operating systems use different forms to specify the path that identifies a particular directory, for example, to set the default directory or to identify the location of a file to be read or written. Within Mathematica, commands for performing these operations should separate directories in a path with the structure that is native to the operating system in use (e.g., forward slash in UNIX, backslash in Windows). For convenience, and to simplify command files to be used with several operating systems, Mathematica will translate forward slashes into the appropriate separator for the operating system in use.

8.18.3 Default Directory

The default directory or current working directory in Mathematica is set when Mathematica is launched. Unless an initialization file has been created with a command that changes that directory, it will probably be the user's home directory or the directory from which Mathematica was launched by a command at a Shell prompt. The command \texttt{Directory[]} will cause the display of the current default directory. As illustrated by the statements

\begin{verbatim}
SetDirectory[]
SetDirectory[".."]
SetDirectory["progs"]
SetDirectory["/apps/MATHEM/userprogs"]
\end{verbatim}

Set directory to user's home directory.
Set directory to the parent of current directory.
Set directory to progs subdirectory of current directory.
Set directory to indicated directory.
the command `SetDirectory` allows dynamic specification of the current working directory in a variety of ways. Note that the separator between successive directories in a path varies from operating system to operating system. The illustrated separator is the one used in UNIX. Further information about these commands will be found in the Mathematica manuals. Further, the statement `FileNames[]` will display a list of the files in the current directory and the statement `FileNames["*.m"]` will display a list of all files in that directory with filetype `.m`. If the file is not in the default directory, then its full path must be specified as part of the filename.

### 8.18.4 Search Path

Whenever an explicit path is not specified for a file that Mathematica is to read, the program looks in a sequence of directories defined by the search path. The initial search path is defined in a Mathematica initialization file or, in the absence of such a file, by Mathematica itself. The initial search path is site-specific and is identified in the Local Guide. If not defined by an initialization file, the path is likely to contain, in order, a few specific Mathematica directories in the user’s home directory, the default directory, the user’s home directory, and finally a succession of directories descending from the directory `$MATHEMHEAD`. The system variable `$Path` contains a list of the directories in the current search path. Thus, the statement `$Path` will display that list, and statements like

```mathematica
Prepend[ $Path, "/usr/share/CPSUP/mathematica" ]
Append[ $Path, "/usr/local/packages" ]
Insert[ $Path, "/usr/local/temp", 5 ]
```

will create a new value for the list `$Path`, that new value containing all of the elements of the original list in the original order but having the indicated directory at the beginning, at the end, or in the fifth position, respectively. Similarly, statements like

```mathematica
Delete[ $Path, 5 ]
ReplacePart[ $Path, "/usr/local/myfiles", 7 ]
```

will create a new value for the list `$Path` by removing altogether the element in the fifth position or by replacing the element in the seventh position with the indicated directory, respectively.

### 8.18.5 Displaying and Changing Default Options

In the course of introducing several commands, we have learned that the precise effect of many commands depends on the values of associated options. Initially, we accepted Mathematica’s defaults. Then we learned that arguments like

`OptionName -> NewValue`

(e.g., `PlotRange -> {-1.0,1.0}`) can be used within a command to override the default value for the indicated option, *but only for that execution of the command*. If the default value is to be overridden for all subsequent executions of the associated command, however, changing the default with a statement like

`SetOptions[ CommandName, OptionName -> NewValue ]`

---

55 See Section 8.18.2.
56 For definiteness, we use the UNIX form for indicating paths. Remember that the precise form in which paths are specified depends on the operating system. The form appropriate at your site is described in the Local Guide.
(e.g., \texttt{SetOptions[ Plot, PlotRange -> \{-1.0,1.0\} ]}) may be more convenient—though this change will, of course, be in effect only until it is changed again (or until a new \textit{Mathematica} session is launched). Finally, the statement

\texttt{Options[ CommandName ]}

(e.g., \texttt{Options[ Plot ]}) will display a list of all options admitted by the indicated command, each with its current default value.

### 8.18.6 Placing Text in Graphs

With suitable commands, we can generate a graphical display to which specified text has been added in the body of the display. Suppose, for example, we wanted to label a graph of the sine function with a string of characters positioned not as a title on the graph as a whole or as a label on an axis but rather as a notation at specified coordinates within the graph. We might use statements like

\begin{verbatim}
In[48]:= Plot[ Sin[x], \{x, 0.0, 10.0\} ];
In[49]:= Graphics[ Text[ "New Graph", \{5.0,0.5\} ] ];
In[50]:= Show[ \{ %48, %49 \} ]
\end{verbatim}

The first of these statements creates a graphics object containing the desired sine curve, the second statement uses the graphics primitive \texttt{Text} to create a second graphics object containing the specified text \textit{centered} at the specified location in the coordinates of the graph and oriented horizontally. And the third statement creates and displays a composite graphics object assembled from

- the graphics object created by line \texttt{In[48]} and known by the name \texttt{\%48}, and
- the graphics object created by line \texttt{In[49]} and known by the name \texttt{\%49}.

We could, of course, have nested the first two statements within the third, but the result would have been syntactically difficult to fathom.

Of course, this quick example only hints at the potential of these commands. The \textit{Mathematica} manuals describe many more graphics primitives that can be used as arguments to the \texttt{Graphics} command and also describe numerous options associated both with the \texttt{Graphics} command and with the \texttt{Show} command. Careful study of those items in the \textit{Mathematica} manuals is prerequisite to full exploitation of \textit{Mathematica}'s abilities to create graphics objects and combine them in elaborate displays.

### 8.18.7 Space Curves

Sometimes it is desirable to view the trajectory of a particle in three-dimensional space. \textsc{Mathematica} has the ability to plot space curves from knowledge of the trajectory defined parametrically by the functions \(x(t), y(t),\) and \(z(t)\) giving the coordinates of points on a three-dimensional path. With that information, the command \texttt{ParametricPlot3D} will calculate the coordinates of the points along the trajectory, project these points onto the two-dimensional screen, and connect consecutive points with lines. To illustrate this feature, consider the equations

\begin{equation}
x(t) = 5 \cos t \quad ; \quad y(t) = 5 \sin t \quad ; \quad z(t) = \alpha t \tag{8.5}
\end{equation}

\footnote{The \texttt{Graphics} command accepts one or two additional arguments, each of which is a list. For example, the alternative command \texttt{Graphics[ Text[ "New Graph", \{5.0,0.5\}, \{-1,0\}, \{0,1\} ]\] will place the left end of the text at the specified point (third argument) and orient the text vertically and reading from the right side of the graph (fourth argument). Other options for the third and fourth arguments are described in the \textit{Mathematica} manuals.}
Figure 8.5: Space curve of a charged particle moving in a constant magnetic field along the $z$ axis.

where $\alpha$ is a constant, describing the trajectory of a charged particle moving in a constant magnetic field directed along the $z$ axis. The starting point is, of course, to evaluate $x$, $y$, and $z$. Then, we invoke the command \texttt{plot3d} to plot the graph. The statements

\begin{verbatim}
In[1]:= alpha = 0.5; \hspace{1cm} \text{Set alpha.}
In[2]:= ParametricPlot3D[ {5*Cos[t], 5*Sin[t], alpha*t}, {t, 0, 30}, \hspace{0.5cm} \text{Invoke ParametricPlot3D to plot trajectory.}
     PlotStyle -> {Thickness[0.02], Black}]
\end{verbatim}

will produce the graph in Fig. 8.5.

\subsection*{8.18.8 Directing Plot to Separate Window}

The several simple graphics commands we have introduced all direct their output to the window in which the commands are issued. Should you wish to have a new graphics window created for a particular plot, you need only add the Option

\texttt{DisplayFunction -> CreateDialog}

to whatever else in incorporated in the command. For example, the statement

\begin{verbatim}
Plot[ Tanh[x], {x, -5.0, 5.0} ]
\end{verbatim}

produces the specified graph in line while the statement

\begin{verbatim}
Plot[ Tanh[x], {x, -5.0, 5.0}, DisplayFunction->CreateDialog ]
\end{verbatim}

directs the graph to a newly created (small) separate window. By the second route, the first graph remains on the screen even if the in line graph scrolls off as further statements are executed.
8.19 References

In this chapter, we have introduced only the most important features of Mathematica. A full description of all commands and features is contained in the Mathematica manuals, including

- several documents and resources as described in Section 8.2.
- a number of documents, links to which can be found by entering “Mathematica help” in the search window of your browser. In particular, you should find links to
  - the official Mathematica documentation at http://reference.wolfram.com/language/, which brings up on line the same documentation available from the HELP menu in the NBI, and
  - many other official and unofficial documents.
- a number of books and on-line materials describing a wide spectrum of uses of Mathematica in a variety of disciplines and for a variety of specific purposes. A list of such books can be found, among other places, from the links http://www.wolfram.com/mathematica/resources/ and http://www.wolfram.com/books/ at the Wolfram Research website or by searching on-line commercial book stores (Amazon, Barnes and Noble, ...) for the keyword Mathematica.
- more specific documents on particular Mathematica commands, links to which can be found by entering “CommandName in MATHEMATICA”, e.g., “integrate in MATHEMATICA” in the search window of your browser.

The user is urged to browse in these materials, being alert not only to the available functions but also to the assorted option variables and keywords that modify their behavior.

While on-line documentation has largely superseded printed documentation in the last decade, several now largely outdated but still potentially valuable printed materials can be found. Specifically, your local library may include one or more copies of

- Getting Started with Mathematica Version # under (operating system), (Wolfram Media, Inc.), which contains installation instructions and a brief description of the issues with which a user of Mathematica must be familiar from the very beginning.
- Stephen Wolfram, The Mathematica Book, Fifth (or earlier) Edition (Wolfram Media, Inc., and Cambridge University Press, 1988, 1991, 1996, 1999, 2003), which contains a tour of the capabilities of Mathematica, an introduction to the practical use of Mathematica, an exhaustive discussion of the design principles underlying Mathematica, a treatment of more advanced topics, several pages of formulae, a section of color plates showing a spectacular array of graphical displays, and a comprehensive set of appendices listing and describing essentially all commands, options, special variables, and other objects that play a role somewhere in Mathematica. This tome of nearly 1500 pages in a single binding is massive and exhaustive.
- Mathematica Version # Standard Add-on Packages (Wolfram Media, Inc., 1996, 1999), which contains detailed descriptions of all components in all of the standard add-on packages.

8.20 Exercises

8.1. Read the Mathematica manuals to find out how the commands Expand, Factor, Apart, and Simplify differ from one another and then invent several examples that will reveal those differences. To help you get started, you might find it useful to compare the effect of each of these commands on the expressions
8.20. **EXERCISES**

\[
\begin{align*}
(a) & \quad \frac{x}{a+b} + \frac{y}{a+b} & \quad (b) & \quad (a+b)(c+d) + (a+b)c \\
(c) & \quad (x+y)(z+w) & \quad (d) & \quad a(b+c(d+e)) \\
(e) & \quad \frac{x}{a+b} + \frac{y}{c+d} & \quad (f) & \quad \frac{(a-x)^2}{(a^2 - 2ax + x^2)^{3/2}}
\end{align*}
\]

but you will probably have to invent other examples as well to reveal as many of the differences as you can. Write two or three paragraphs in which you describe the action of each of these commands and the differences among them. Creating a table showing the results of applying each of the commands to each expression you select might be a useful support for your comparison.

8.2. Use **Mathematica** to convert each of the expressions in the left-hand column in the table below into the expression in the associated right-hand column:

\[
\begin{align*}
(a) & \quad \frac{(a-x)^2}{(a^2 - 2ax + x^2)^{3/2}} \quad \Rightarrow \quad \frac{1}{|x-a|} \\
(b) & \quad \sinh(\ln(x + \sqrt{x^2 + a^2}) - \ln(a)) \quad \Rightarrow \quad \frac{x}{a} \\
(c) & \quad \frac{1}{x + \sqrt{y}} \quad \Rightarrow \quad \frac{x - \sqrt{y}}{x^2 - y}
\end{align*}
\]

Write two or three paragraphs in which you describe your efforts, including some indication of approaches that were not successful. Don’t be overly concerned about the order of terms within various sets of parentheses; that order is particularly difficult to control. Focus instead on creating the general form of each desired result.

8.3. Use **Mathematica** to convert each of the expressions in the left-hand column in the table below into the expression in the associated right-hand column:

\[
\begin{align*}
(a) & \quad \frac{d}{dx} \left( x e^{-x^2} \right) \quad \Rightarrow \quad -2x(x^2 - 1)e^{-x^2} \\
(b) & \quad \sin(\sqrt{a^2(a+3x) + x^2(3a+x) + y}) \quad \Rightarrow \quad \sin((a + x)^{3/2} + y) \\
(c) & \quad cg + cf + b^2d + 2abcd + a^2d + b^2c + 2abc + a^2c \quad \Rightarrow \quad (a + b)^2(c + d) + c(f + g) \\
(d) & \quad a e^{(-b+i\omega)t} + a e^{(-b-i\omega)t} \quad \Rightarrow \quad a e^{-bt} (e^{-i\omega t} + e^{i\omega t}) \\
(e) & \quad a e^{(-b+i\omega)t} + a e^{(-b-i\omega)t} \quad \Rightarrow \quad 2a e^{-bt} \cos(\omega t) \\
(f) & \quad x^2 + y^2 + z^2 - 2a(x + y) + 2a^2 \quad \Rightarrow \quad (x - a)^2 + (y - a)^2 + z^2
\end{align*}
\]

In several cases, you may need to invoke **Part, ReplacePart**, and/or **FullForm**, but other commands will surely also be needed. Invoke microscopic dissection of the expressions only as a last resort. Write two or three paragraphs in which you describe your efforts, including some indication of approaches that were not successful. Don’t be overly concerned about the order of terms within various sets of parentheses in the final form; that order is particularly difficult to control. Focus instead on creating the general form of each desired result.

8.4. In Section 8.8.9, we set up the differential equation for a damped harmonic oscillator but then pursued the solution only for the underdamped case. Starting with the dimensionless form\(^{58}\)

\[
\frac{d^2 x}{dt^2} + 2\alpha \frac{dx}{dt} + x = 0 \quad ; \quad x(0) = 1 \quad ; \quad \frac{dx}{dt} \bigg|_{t=0} = 0
\]

\(^{58}\)The forms of the solution work out more easily if, in contrast to the choice in the text, we express the coefficient of the second term as twice a constant, so \(\alpha = b/2m\omega_0\).
of the equation deduced towards the end of Section 8.8.9, work out the solution for the underdamped case and then find also the solutions for the critically damped and overdamped cases defined, respectively by

$$\alpha > 0 \quad \text{and} \quad \alpha <, =, > 1$$

Describe the differences in the physical behavior for the three cases. For simplicity, take the initial conditions to be $x(0) = 1$ and $v(0) = 0$.

8.5. The Legendre polynomials $P_n(x)$, which are valid and useful over the interval $-1 \leq x \leq 1$, can be defined in many ways. They emerge as the coefficients in the Taylor expansion of the generating function

$$g(x, t) = \frac{1}{\sqrt{1 - 2tx + t^2}} = \sum_{n=0}^{\infty} P_n(x) t^n$$

Alternatively, they can be determined from the recursion relationship

$$(2n + 1) x P_n(x) = (n + 1) P_{n+1}(x) + n P_{n-1}(x)$$

provided we include the first two $P_0(x) = 1$ and $P_1(x) = x$ to get started. Yet again, they can be found from application of multiple differentiation as implied by Rodrigues’ formula

$$P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} \left((x^2 - 1)^n\right)$$

However they are determined, the first half dozen of these polynomials will turn out to be

$$P_0(x) = 1 \quad P_2(x) = \frac{1}{2} (3x^2 - 1) \quad P_3(x) = \frac{1}{2} (5x^3 - 3x)$$
$$P_1(x) = x \quad P_4(x) = \frac{1}{8} (35x^4 - 30x^2 + 3)$$
$$P_5(x) = \frac{1}{8} (63x^5 - 70x^3 + 15x)$$

a. Use the generating function and Mathematica’s capabilities for evaluating Taylor series to find the first half-dozen Legendre polynomials, extracting each as an expression bound to a variable. The commands Collect and CoefficientList may be useful.

b. Start by binding the value 1 to $P[0]$ and the value $x$ to $P[1]$. Then, using the recursion relationship, find the next several Legendre polynomials. Hint: The Mathematica statements might be

In[29]:= P[0] = 1; P[1] = x;
and then

In[30]:= PP = (2*n-1)*x*P[n-1]/n - (n-1)*P[n-2]/n

(Verify the expression on the right by using Mathematica to deduce this relationship from the standard form—the second equation in this exercise.) With these statements, you have set up $P_0(x)$ and $P_1(x)$ to start the recursion and then you have defined an expression $PP$ depending on $n$ that can be evaluated at any $n$. Once $P_0$ and $P_1$ have been defined, you can find $P_2$ and then $P_3$ and then ... with statements like

In[31]:= P[2] = PP //. n -> 2
In[32]:= P[3] = PP //. n -> 3

...  

c. Find the first half-dozen Legendre polynomials by using the command D to evaluate Rodrigues’ formula. Hint: You might find that using a loop would simplify your approach.

d. Be clever and, using either matrices or loops constructed in Mathematica, find the values of all of the integrals

$$\int_{-1}^{1} P_n(x) P_m(x) \, dx$$

where $n$ and $m$ take on independently the values 0, 1, 2, 3, 4, 5. (There are 36 integrals to be evaluated. Try to be efficient in your coding, and note that the command Integrate applied to a matrix will automatically apply the command element-by-element.)
e. Within Mathematica, obtain graphs of the first six Legendre polynomials over the interval $-1 \leq x \leq 1$.

f. It is known that a function $f(x)$ defined over the interval $-1 \leq x \leq 1$ can be expanded in a series of Legendre polynomials of the form $f(x) = \sum a_n P_n(x)$ where the coefficients $a_n$ are given by

$$a_n = \frac{2n + 1}{2} \int_{-1}^{1} f(x) P_n(x) \, dx$$

Find the first six coefficients for the expansion of the function $f(x) = 0$ when $-1 < x < 0$ and $f(x) = 1$ when $0 < x < 1$. Then, construct the (partial) series representing this function and obtain a graph of that approximation to compare with the graph of the original function.

g. Mathematica actually knows quite a bit about many of the important special functions of mathematical physics. In particular, the $n$-th Legendre polynomial as a function of $x$ is known by the name `LegendreP[n, x]` and is described fully in the Mathematica manuals. Take a look at that documentation and then use that function to determine the first several Legendre polynomials.

8.6. Figure 8.6 shows the circuit diagram for a Wheatstone bridge. Using Kirchoff’s laws, set up the equations from which you could determine the currents in each branch of the circuit. Then, using Mathematica, (a) solve the equations symbolically, (b) find conditions under which the current in the cross branch (through resistor $R_5$) will be zero, and (c) find the effective resistance seen by the battery. (The effective resistance is defined by the ratio $V/I$, where $I$ is the current in the branch containing the battery.)

8.7. In the scattering of a quantum wave from a rectangular barrier in particular circumstances, we find that the wave function must be expressed in three pieces in the form

$$
\psi(x) = \begin{cases} 
Ae^{ikx} + Be^{-ikx} & x < 0 \\
D\cosh(\kappa x) + F\sinh(\kappa x) & 0 < x < w \\
Ce^{ikx} & x > w
\end{cases}
$$

where $k$ and $\kappa$ are constants related to the energy of the particle and the height of the barrier, $A$ is a constant reflecting the intensity of the incident beam, and $B$, $D$, $F$, and $C$ are constants to be determined by imposing the requirement that the wave function and its first derivative be continuous both at $x = 0$ and at $x = w$, i.e., that

$$
\psi(0^-) = \psi(0^+) ; \quad \psi(w^-) = \psi(w^+) ; \quad \frac{d\psi}{dx}\bigg|_{0-} = \frac{d\psi}{dx}\bigg|_{0+} ; \quad \frac{d\psi}{dx}\bigg|_{w-} = \frac{d\psi}{dx}\bigg|_{w+}
$$

where superscript plus and minus signs identify points slightly below and slightly above the indicated value of $x$, respectively. (While $k$ and $\kappa$ can be taken to be real for this barrier, the constants $A$, $B$, $C$, $D$, and $F$ may—and probably will—be complex.) Use Mathematica’s abilities to manipulate expressions to
(a) Obtain the equations determining $B$, $C$, $D$, and $F$ by imposing the stated boundary conditions on these solutions.

(b) Solve those equations for those constants (expressing each as a multiple of the constant $A$).

(c) Show that the reflection and transmission coefficients $R$ and $T$ defined by $R = |B/A|^2$ and $T = |C/A|^2$ are given by

$$R = \left| \frac{B}{A} \right|^2 = \frac{(k^2 + k^2)^2 \sinh^2 \kappa \omega}{4k^2(k^2 + \kappa^2)^2 \sinh^2 \kappa \omega}; \quad T = \left| \frac{C}{A} \right|^2 = \frac{4k^2 \kappa^2}{4k^2(k^2 + \kappa^2)^2 \sinh^2 \kappa \omega}$$

Here, the vertical bars symbolize the absolute value of the complex number enclosed by them.

(d) Verify that $R + T = 1$.

8.8. The distribution of wavelengths $\lambda$ in the blackbody spectrum at (absolute) temperature $T$ is given by

$$u(\lambda, T) = \frac{8\pi c h \lambda^3}{\exp\left(\frac{h c}{\lambda k T}\right) - 1}$$

where $c$ is the speed of light, $h$ is Planck’s constant, and $k$ is Boltzmann’s constant. In terms of the variable $y = \frac{ch}{\lambda k T}$, this function has the alternative expression

$$\frac{(ch)^4 u(\lambda, T)}{8\pi (kT)^5} = f(y) = \frac{y^5}{e^y - 1}$$

(a) Verify this transformed form and then, using Mathematica, (b) obtain a graph of $f(y)$ versus $y$, making sure to extend the graph over an interval that includes its peak and estimate the value of $y$ at which that peak occurs; (c) show that the peak occurs for values of $y$ satisfying

$$(y - 5) e^y + 5 = 0$$

(d) obtain a graph of this function versus $y$; (e) make another estimate of the value of $y$ at which the original function has its maximum; and (f) show that the wavelength $\lambda_m$ at which this maximum occurs satisfies

$$\lambda_m T = 0.28978 \times 10^{-2} \text{ m K}$$

Hints: (1) Remember that maxima in a function occur where the derivative of that function with respect to the appropriate variable is zero. (2) Because $e^y$ varies rapidly with $y$, you may have to play a bit to find a suitable range of values of $y$ over which to plot these graphs.

8.9. Use Mathematica to verify (a) that the solution

$$x(t) = e^{-t^a} x_0 \left( \cos(at) + \frac{\gamma \sin(at)}{a} \right)$$

with

$$a = \sqrt{\omega^2 - \gamma^2} \quad ; \quad \gamma = \frac{b}{2m} \quad ; \quad \omega^2 = \frac{k}{m}$$

presented in Section 8.8.9 satisfies the desired differential equation and the initial conditions presented at the beginning of that section and (b) that this solution and the solution

$$y(\tau) = \frac{x_0}{2\sqrt{1 + \beta^2}} \left( -e^{(-\beta - \sqrt{1 + \beta^2})\tau} + e^{(-\beta + \sqrt{1 + \beta^2})\tau} \beta + e^{(-\beta - \sqrt{1 + \beta^2})\tau} \sqrt{1 + \beta^2} + e^{(-\beta + \sqrt{1 + \beta^2})\tau} \sqrt{1 + \beta^2} \right)$$

where $\tau = \omega t$ and $\beta = \gamma/\omega$ obtained at the very end of the section agree with each other.

8.10. The Laguerre polynomials $L_n(x)$ can be defined in many ways. We might, for example, set $L_0(x) = 1$ and then generate each new polynomial in turn by requiring that $L_n(x)$ for $n > 0$ be a polynomial of order $n$ that is orthogonal to all previous polynomials, i.e., that

$$\int_0^\infty e^{-x} L_n(x) L_m(x) \, dx = 0 \quad , \quad \text{all } m < n$$
and that the arbitrary overall sign remaining be resolved by requiring in addition that \( L_n(0) = 1 \).
Alternatively, they can be determined from the recursion relationship

\[
(n + 1)L_{n+1}(x) = (2n + 1 - x)L_n(x) - nL_{n-1}(x)
\]

provided we include the first two \( L_0(x) = 1 \) and \( L_1(x) = 1 - x \) to get the recursion started. Yet again, they can be found from the application of multiple differentiation as implied by Rodrigues’ formula

\[
L_n(x) = \frac{e^x}{n!} \frac{d^n}{dx^n}(x^n e^{-x})
\]

However they are determined, the first half dozen of these polynomials will turn out to be

\[
\begin{align*}
L_0(x) &= 1 \\
L_1(x) &= 1 - x \\
L_2(x) &= 1 - 2x + \frac{1}{2}x^2 \\
L_3(x) &= 1 - 3x + \frac{3}{2}x^2 - \frac{1}{6}x^3 \\
L_4(x) &= 1 - 4x + 3x^2 - \frac{2}{3}x^3 + \frac{1}{3}x^4 \\
L_5(x) &= 1 - 5x + 5x^2 - \frac{5}{3}x^3 + \frac{5}{6}x^4 - \frac{1}{120}x^5
\end{align*}
\]

a. Determine the first half dozen Laguerre polynomials by constructing them one at a time to satisfy the requirements of orthogonality and normalization.

b. Start by binding the value 1 to \( L[0] \) and the value \( 1 - x \) to \( L[1] \). Then, using the recursion relationship, find the next several Laguerre polynomials. **Hint:** The _Mathematica_ statements might be

\[
\begin{align*}
\text{In}[29]:= & \ L[0] = 1; \ L[1] = 1-x; \\
\text{In}[30]:= & \ \text{LL} = (2*n-1-x)*L[n-1]/n - (n-1)*L[n-2]/n
\end{align*}
\]

(Verify the expression on the right.) With these statements, you have set up \( L_0(x) \) and \( L_1(x) \) to start the recursion and then you have defined an expression \( \text{LL} \) involving \( n \) that can be evaluated at any \( n \). Once \( L_0 \) and \( L_1 \) have been defined, you can find \( L_2 \) and then \( L_3 \) and then \( L_4 \) and then \( L_5 \) with statements like

\[
\begin{align*}
\text{In}[31]:= & \ L[2] = \text{LL} /. \ n \rightarrow 2 \\
\text{In}[32]:= & \ L[3] = \text{LL} /. \ n \rightarrow 3 \\
\cdots
\end{align*}
\]

c. Find the first half-dozen Laguerre polynomials by using the command \( \text{D} \) to evaluate Rodrigues’ formula.

d. Be clever and, using either matrices or loops constructed in MATHEMATICA, find the values of all of the integrals

\[
\int_0^\infty e^{-x} L_n(x) L_m(x) \, dx
\]

where \( n \) and \( m \) take on independently the values 0, 1, 2, 3, 4, 5. (There are 36 integrals to be evaluated. Try to be efficient in your coding, and remember the command \text{map}, which will be useful in constructing a single statement that evaluates the integral of each element in a multi-element structure.)

e. _Mathematica_ actually knows quite a bit about many of the important special functions of mathematical physics. In particular, the \( n \)-th Laguerre polynomial as a function of \( x \) is known by the name \text{LaguerreL}[n,x] \] and is described fully in the _Mathematica_ manuals. Take a look at that documentation and then use that function to determine the first several Laguerre polynomials.

8.11. Find the eigenvalues and eigenvectors of the matrix

\[
\begin{pmatrix}
0 & 1 & 0 \\
1 & 0 & 1 \\
0 & 1 & 0
\end{pmatrix}
\]
8.12. (a) Use the function \texttt{Eigensystem} to find the eigenvalues and eigenvectors of the matrix
\[
\begin{pmatrix}
2 & -1 & 0 & 0 \\
-1 & 2 & -1 & 0 \\
0 & -1 & 2 & -1 \\
0 & 0 & -1 & 2
\end{pmatrix}
\]

(b) Use the function \texttt{FullForm} to explore the structure of the entity returned by the function \texttt{Eigensystem} used in part (a).

8.13. When a (weak) constant external electric field of magnitude $F$—we reserve $E$ for energy in this exercise—is imposed on a hydrogen atom, the energy of the states with principal quantum number $n$ shift from the energy given by the Bohr model by amounts determined by the eigenvalues of the matrix whose elements are $\langle nlm|eF\hat{z}|nl'm'\rangle$, where $l$, $m$, $l'$, and $m'$ range over all possible values of those quantum numbers allowed by the particular value of $n$. If the states by which the rows and columns are labeled are ordered $|2,0,0\rangle$, $|2,1,-1\rangle$, $|2,1,0\rangle$, and $|2,1,1\rangle$, then the matrix for the state $n = 2$ is
\[
3eaoF \begin{pmatrix}
0 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 \\
-1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}
\]
where $e$ is the magnitude of the charge on the electron and $a_0$ is the Bohr radius. Similarly, if the states by which the rows and columns are labeled are ordered $|3,2,2\rangle$, $|3,1,1\rangle$, $|3,2,1\rangle$, $|3,0,0\rangle$, $|3,1,0\rangle$, $|3,2,0\rangle$, $|3,1,-1\rangle$, $|3,2,-1\rangle$, and $|3,2,-2\rangle$, then the matrix for the state $n = 3$ is
\[
3eaoF \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -9/2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -9/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -3\sqrt{6} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -3\sqrt{6} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -9/\sqrt{3} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -9/\sqrt{3} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -9/2 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]
Find the eigenvalues and eigenvectors of these matrices. The eigenvalues give the energy shifts for the Stark effect for $n = 2$ and $n = 3$ and the eigenvectors give the linear combinations of the base states (i.e., the states in the absence of the external field) out of which the states in the presence of the field emerge as the field is turned on. \textit{Hint}: The simplest way to create a sparse matrix is to begin by executing the statements
\[
\texttt{<< LinearAlgebra'\texttt{MatrixManipulationelianAlgebra'\texttt{MatrixManipulation}}\texttt{'}} \\
\texttt{mat = ZeroMatrix[4,4];}
\]
which load the add-on package that defines the command \texttt{ZeroMatrix} and create a $4 \times 4$ matrix of zeros. Continuing, we then set the nonzero entries with statements like
\[
\texttt{mat[[1,3]]} = \text{desired value}
\]
For verification, the final matrix can be displayed in matrix form with the statement
\[
\texttt{MatrixForm[ mat ]}
\]

8.14. As shown in Fig. 8.7, an object of mass $m$ is connected to the center of each of the four sides of a square of sides $2\ell$ with a spring of constant $k$ and moves on a horizontal frictionless surface in the plane defined by the square. Take the equilibrium position of the object at the center of the square to be the origin of an $xy$ coordinate system, and let the springs have unstretched length $\ell_0$. For this situation, the kinetic and potential energies $KE$ and $PE$ of the object are given by
\[
KE = \frac{1}{2}mx^2 + \frac{1}{2}my^2 ; \quad PE = \frac{1}{2}k\left[\sqrt{(\ell - x)^2 + y^2 - \ell_0}\right] + \ldots
\]
where $PE$ will have three additional terms, one for each of the remaining springs. (The term shown applies to the spring connected to the right side of the square.) By definition, the Lagrangian function for this problem is given by $L = KE - PE$ and the equations of motion can be extracted from the Lagrangian by evaluating the expression

$$\frac{d}{dt} \left( \frac{\partial L}{\partial \dot{q}} \right) - \frac{\partial L}{\partial q} = 0$$

where $q$ stands first for $x$ and then for $y$. Write out the full potential energy and then use Mathematica to deduce the equations of motion $m\ddot{x} = \ldots$ and $m\ddot{y} = \ldots$ for this system. The results are quite complicated. To simplify the problem, find in detail the equations only for the cases (a) $\ell_0 = 0$, i.e., the springs have unstretched length of 0—admittedly unrealistic springs, and (b) $x$ and $y$ remain small compared to $\ell$ throughout the motion, i.e., $x/\ell \ll 1$ and $y/\ell \ll 1$ for all time. 

**Note:** Statements like $D[L, D[x[t], t]]$ are perfectly acceptable to Mathematica, provided the dependence of $x$ on $t$ is explicitly indicated in setting up the Lagrangian $L$. Thus, for example, in a different problem (particle of mass $m$ falling under the gravitational attraction of the earth), for which we might set

$$L = m\times D[x[t], t]^2/2 - m\times g\times x[t]$$

we can evaluate $\partial L/\partial \dot{x}$ with the statement

```
DLDx = D[ L, x[t] ]
```

evaluate $\partial L/\partial x$ with the statement

```
DLDx = D[ L, x[t] ]
```

and then, finally, determine the equation of motion with the statement

```
D[ DLDx, t] - DLDx == 0
```

8.15. In the Cartesian coordinate system illustrated in Fig. 8.8, the coordinates of the two objects having mass $m_1$ and $m_2$ in the compound pendulum are given by

$$x_1 = l_1 \sin \theta; \quad y_1 = -l_1 \cos \theta; \quad x_2 = l_1 \sin \theta + l_2 \sin \phi; \quad y_2 = -l_1 \cos \theta - l_2 \cos \phi$$

Taking $\theta$ and $\phi$ to be the generalized coordinates and remembering that

$$KE = \frac{1}{2}m_1 x_1^2 + \frac{1}{2}m_1 y_1^2 + \frac{1}{2}m_2 x_2^2 + \frac{1}{2}m_2 y_2^2; \quad PE = m_1 g y_1 + m_2 g y_2$$

use Mathematica to find (a) (simple) expressions for $KE$, $PE$, and $L = KE - PE$, (b) the equations of motion by evaluating the Lagrange equations

$$\frac{d}{dt} \left( \frac{\partial L}{\partial \dot{\theta}} \right) - \frac{\partial L}{\partial \theta} = 0; \quad \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{\phi}} \right) - \frac{\partial L}{\partial \phi} = 0$$
and (c) the equations of motion when both $\theta$ and $\phi$ are small. Suggestion: If you run into major difficulties keeping $l_1 \neq l_2$ and $m_1 \neq m_2$, make the lengths and masses the same. Note: See the note in the previous exercise.

8.16. The complete elliptic integrals of the first and second kinds are given by

$$K(k) = \int_0^{\pi/2} \frac{d\phi}{(1 - k^2 \sin^2 \phi)^{1/2}} ; \quad E(k) = \int_0^{\pi/2} (1 - k^2 \sin^2 \phi)^{1/2} d\phi$$

Use Mathematica to evaluate these integrals to $O(k^6)$ by expanding the integrands in Taylor series before evaluating the integrals.

8.17. Find the Laplace transform of each of the functions

(a) $f(t) = t^n$ (b) $f(t) = t e^{-at}$ (c) $f(t) = \cosh(at)$ (d) $\frac{df(t)}{dt}$

and the inverse Laplace transform of each of the functions

(e) $\tilde{f}(s) = \frac{a + bs}{s^2 + \omega^2}$ (f) $\tilde{f}(s) = \frac{a}{(s^2 + 9\omega^2)(s^2 + 4\omega^2)(s^2 + \omega^2)}$

In (c), you may have to express the hyperbolic cosine in exponential form while in (f) you may have to help Mathematica’s routine InverseLaplaceTransform by first invoking the command Apart to expand the desired function in partial fractions.

8.18. Given the three points $(x_i, y_i), i = 1, 2, 3$, (a) find symbolic expressions for the coefficients $a$, $b$, and $c$ of the parabola $y = ax^2 + bx + c$ that passes through these three points and then (b) find a symbolic expression for the value of $x$ at which the extreme point of the parabola occurs. Finally, (c) determine numerically the angle at which the maximum range of a projectile occurs if the ranges at $\theta = 39^\circ$, $40^\circ$, and $41^\circ$ are 0.7251744, 0.7259383, and 0.7258887, respectively.

8.19. Find symbolic expressions for the coefficients $a$, $b$, and $c$ that will cause the parabola $y = ax^2 + bx + c$ to pass through the three points $(x_1 = x_2 - \delta x, y_1)$, $(x_2, y_2)$, and $(x_3 = x_2 + \delta x, y_3)$. Then integrate the parabola over the interval $x_1 \leq x \leq x_3$ and deduce Simpson’s rule

$$\int_{x_1}^{x_3} y(x) \, dx = \frac{\delta x}{3} \left( y_1 + 4y_2 + y_3 \right)$$

8.20. The midpoint rule $M$ and Simpson’s rule $S$ for evaluating integrals numerically start with the assumptions that

$$\int_a^b f(x) \, dx \approx M = (b - a) f \left( \frac{a + b}{2} \right) ; \quad \int_a^b f(x) \, dx \approx S = \frac{b - a}{6} \left( f(a) + 4f \left( \frac{a + b}{2} \right) + f(b) \right)$$
8.20. EXERCISES

8.21. Using Mathematica, obtain graphs of the potential

\[ U(x) = -\frac{U_0 a^2 (a^2 + x^2)}{8a^4 + x^4} \]

and the associated force. Note that the expression for the potential is simpler to plot if you plot not \( U(x) \) versus \( x \) but rather \( U(x)/U_0 \) versus \( x/a \), i.e., rewrite the function in the form

\[ \frac{U(x)}{U_0} = -\frac{1 + x^2}{8 + x^4} \]

You should be able to find a similar dimensionless version of the expression giving the force.

8.22. Using Mathematica, find complete (symbolic) solutions to each of the following problems, and use Mathematica also to verify that the solutions you obtain actually do satisfy the original ODE and initial conditions.

a. \( \frac{d^2 x}{dt^2} = a \), \( x(0) = x_0 \), \( v(0) = v_0 \), where \( a \) is constant, i.e., find position as a function of time for a particle moving under the action of a constant force and launched with arbitrary initial conditions.

b. \( m \frac{d^2 x}{dt^2} = -eE_0 \cos(\omega t + \theta) \), \( x(0) = x_0 \), \( v(0) = v_0 \), i.e., find position as a function of time for a charged particle moving under the action of a sinusoidal force and launched with arbitrary initial conditions.

c. \( m \frac{d^2 x}{dt^2} = -mg + b \left( \frac{dx}{dt} \right)^2 \), \( x(0) = 0 \), \( v(0) = 0 \), i.e., find position as a function of time for a particle released from rest at the origin and allowed to fall freely under the action of gravity and a viscous retarding force proportional to the square of the speed.

d. the differential equations

\[ m \frac{d^2 x}{dt^2} = -b \frac{dx}{dt} \quad \text{and} \quad m \frac{d^2 z}{dt^2} = -mg - b \frac{dz}{dt} \]

for the motion of a projectile moving under gravity in a viscous medium when the motion starts at the origin with the initial velocity \( \mathbf{v} = v_0 \hat{i} + v_0 \hat{k} \).

8.23. Using the functions Grad, Laplacian, Div, and Curl defined in Section 8.8.10,

a. Evaluate the gradient of (i.e., the negative of the force field associated with) each of the functions

\[ V_1(x, y, z) = \frac{1}{(x^2 + y^2 + z^2)^{1/2}} \quad V_2(x, y, z) = \frac{z}{(x^2 + y^2 + z^2)^{3/2}} \]

\[ V_3(x, y, z) = \frac{2z^2 - x^2 - y^2}{(x^2 + y^2 + z^2)^{5/2}} \quad V_4(x, y, z) = \frac{e^{-a(x^2+y^2+z^2)^{1/2}}}{(x^2 + y^2 + z^2)^{3/2}} \]

of the Cartesian variables \((x, y, z)\).

b. Verify that each field obtained in part (a) is conservative by showing that the curl of each is zero.
c. With \( \mathbf{r} = x \hat{i} + y \hat{j} + z \hat{k} \), show (1) that \( \nabla \times \mathbf{r} = 0 \) and (2) that \( \nabla \cdot \mathbf{r} = 3 \).
d. With \( V_5(x, y, z) = x^2 + y^2 + z^2 \), show that \( \nabla^2 V_5(x, y, z) = \nabla \cdot \nabla V_5(x, y, z) = 6 \).
e. Show that \( \nabla^2 V_1(x, y, z) = 0 \). (Note: Except at \( x = y = z = 0 \))
f. Evaluate the Laplacian of \( V_4(x, y, z) \), \( \nabla^2 V_4(x, y, z) \).

8.24. Using the functions \texttt{Grad}, \texttt{Laplacian}, \texttt{Div}, and \texttt{Curl} defined in Section 8.8.10,

a. Evaluate the gradient of (i.e., the negative of the force field associated with) each of the functions

\[
V_1(r, \phi, z) = \frac{1}{(r^2 + z^2)^{1/2}} \quad V_2(r, \phi, z) = \frac{z}{(r^2 + z^2)^{3/2}}
\]

\[
V_3(r, \phi, z) = \frac{2z^2 - r^2}{(r^2 + z^2)^{3/2}} \quad V_4(r, \phi, z) = \frac{e^{-a(r^2 + z^2)^{1/2}}}{(r^2 + z^2)^{5/2}}
\]

of the cylindrical variables \( r \) (radial), \( \phi \) (azimuthal), and \( z \) (axial).
b. Verify that each field obtained in part (a) is conservative by showing that the curl of each is zero.
c. With \( \mathbf{r} = r \hat{r} + z \hat{k} \), show (1) that \( \nabla \times \mathbf{r} = 0 \) and (2) that \( \nabla \cdot \mathbf{r} = 3 \).
d. With \( V_5(r, \phi, z) = r^2 + z^2 \), show that \( \nabla^2 V_5(r, \phi, z) = \nabla \cdot \nabla V_5(r, \phi, z) = 6 \).
e. Show that \( \nabla^2 V_1(r, \phi, z) = 0 \). (Note: Except at \( r = z = 0 \))
f. Evaluate the Laplacian of \( V_4(r, \phi, z) \), \( \nabla^2 V_4(r, \phi, z) \).

8.25. Using the functions \texttt{Grad}, \texttt{Laplacian}, \texttt{Div}, and \texttt{Curl} defined in Section 8.8.10,

a. Evaluate the gradient of (i.e., the negative of the force field associated with) each of the functions

\[
V_1(r, \theta, \phi) = \frac{1}{r} \quad V_2(r, \theta, \phi) = \frac{z}{r^3} \quad V_3(r, \theta, \phi) = \frac{r^2(3 \cos^2 \theta - 1)}{r^5} \quad V_4(r, \theta, \phi) = \frac{e^{-ar^2/2}}{r}
\]

of the spherical coordinates \( r \) (radial), \( \theta \) (polar), and \( \phi \) (azimuthal).
b. Verify that each field obtained in part (a) is conservative by showing that the curl of each is zero.
c. With \( \mathbf{r} = r \hat{r} \), show (1) that \( \nabla \times \mathbf{r} = 0 \) and (2) that \( \nabla \cdot \mathbf{r} = 3 \).
d. With \( V_5(r, \theta, \phi) = r^2 \), show that \( \nabla^2 V_5(r, \theta, \phi) = \nabla \cdot \nabla V_5(r, \phi, z) = 6 \).
e. Show that \( \nabla^2 V_1(r, \theta, \phi) = 0 \). (Note: Except at \( r = z = 0 \))
f. Evaluate the Laplacian of \( V_4(r, \theta, \phi) \), \( \nabla^2 V_4(r, \theta, \phi) \).

8.26. Create four three-component vectors \( \mathbf{A}, \mathbf{B}, \mathbf{C}, \) and \( \mathbf{D} \) with statements like

\[
\mathbf{A} = \{A_1, A_2, A_3\}; \quad \mathbf{B} = \{B_1, B_2, B_3\}; \quad \mathbf{C} = \{C_1, C_2, C_3\}; \quad \mathbf{D} = \{D_1, D_2, D_3\};
\]

Then, using the file \texttt{crossdot.m} as described in Section 8.11, show that

a. \( \mathbf{A} \times \mathbf{B} = -\mathbf{B} \times \mathbf{A} \), or, equivalently, that \( \mathbf{A} \times \mathbf{B} + \mathbf{B} \times \mathbf{A} = 0 \).
b. \( \mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = (\mathbf{A} \cdot \mathbf{C}) \mathbf{B} - (\mathbf{A} \cdot \mathbf{B}) \mathbf{C} \)
c. \( (\mathbf{A} \times \mathbf{B}) \times \mathbf{C} = (\mathbf{A} \cdot \mathbf{C}) \mathbf{B} - (\mathbf{B} \cdot \mathbf{C}) \mathbf{A} \)
d. \( \mathbf{A} \cdot (\mathbf{B} \times \mathbf{C}) = (\mathbf{A} \times \mathbf{B}) \cdot \mathbf{C} \)
e. \( (\mathbf{A} \times \mathbf{B}) \cdot (\mathbf{C} \times \mathbf{D}) = (\mathbf{A} \cdot \mathbf{C})(\mathbf{B} \cdot \mathbf{D}) - (\mathbf{A} \cdot \mathbf{D})(\mathbf{B} \cdot \mathbf{C}) \)

59 The variables \( \mathbf{C} \) and \( \mathbf{D} \) in MATHEMATICA are protected.
Chapter 9

Introduction to Programming

In this chapter, we address the task of composing sets of statements\(^1\) that will cause an obedient, efficient, and very literal servant (i.e., computer) to perform exactly the task intended by the master (i.e., programmer). A general strategy by which a particular task can be accomplished in a finite (though perhaps large) number of steps is called an algorithm. A specific expression of an algorithm in a suitable language is called a program. Hence, this chapter is about designing algorithms and implementing them in programs. More specifically, it is about designing algorithms for the performance of tasks that will ultimately be assigned to a computer and about implementing those algorithms in several possible computer languages.

Designing algorithms and implementing them in programs that direct a computer to perform various tasks is really not very different from designing algorithms and implementing them in programs that direct a baker to bake a cake, a knitter to knit a sweater, or a cab driver to drive from the airport to the hotel. In the aggregate, an algorithm will obtain all necessary inputs, manipulate those inputs in some way, and produce the required outputs. Each step in the process must be systematically specified using elementary statements, each of which means only one thing to the servant (person or computer) that will perform the task. Some of the appropriate elementary statements—particularly the action statements that specify simple actions—will vary with the general type of task to be performed. “Mix” and “whip”, for example, are among the elementary action statements that must be understood by the baker of cakes; “knit”, “purl”, and “cast on” must be part of the vocabulary of a knitter of sweaters; while “turn right” and “follow interstate 90 west” are action statements for the cab driver.

The performance of even simple tasks, however, entails the execution of suitable action statements in the right order. The complete description of an algorithm must therefore indicate not only which actions are to be performed but also the order of their performance. Those aspects of an algorithm that specify the order of performance of action statements are called control structures.

A language for specifying algorithms must therefore provide not only elementary action statements, which vary with the general type of task to be performed, but also control structures, which are more universal than the action statements. In the first two sections of this chapter, we focus particularly on identifying fundamental control structures and illustrating their role in several simple algorithms. In later sections, we explain how the general structures introduced in the first section are implemented in standard languages. More detailed information about various languages may be found in the references listed in Section 9.17.

\(^1\)Individual instructions in a computer language are usually called statements, and we use that word throughout this book.
9.1 Components of a Programming Language

While we could code programs directly in the binary language that computers use internally, the process would be slow, tedious, prone to error, and resistant to debugging. As an alternative, a wide assortment of high-level languages has been developed. Using a chosen high-level language, the programmer creates an ASCII file containing the source code, which lays out what the computer must do to accomplish the desired task. Then, that source code must be translated into machine code—a task that a computer equipped with the proper translating program can do for itself.

Actually, there are two approaches to this translation. In the first, the computer interprets the source code, which means that each statement in the code is translated by the computer each time the statement is executed; execution is slow but the process for us is simple: we write the source code and run the program. In the second approach, the computer first compiles the source code into binary object code, and then links the object code with an assortment of system routines and possibly additional user-written routines. Only after compilation and linking have been completed can the resulting binary or executable file actually be run. This second approach is more complicated for us, since we must both write the source code and then compile and link it. The advantages are twofold: (1) the final executable file is stored in the machine and can be run any number of times without repeating compilation and linking; and (2) execution of compiled code is faster than execution when the code is interpreted.

The prospective user of a computer need only learn how to construct the source code in whatever language is to be used. While the detailed syntax of the statements that can be constructed depends very much on the language chosen, certain common elements can be identified. In particular, all languages provide ways to make use of computer memory, to obtain input and display output, and to control the flow of execution within the program. In this section, we describe those common elements as a prelude to laying out the details for specific languages in later sections.

9.1.1 Variables, Variable Names, and Use of Memory

Computers provide a means for working with data. The data, however, must be accessible to the central processing unit (CPU), which is the part of the hardware that directs and controls all actions carried out by the computer. Let us, therefore, endow the CPU with a capacity to reserve an available memory cell (or a contiguous succession of them if one cell is not large enough) and assign to it (them) a variable name, which we specify. Throughout the execution of the program, the assigned name then refers to this unique cubbyhole in the machine’s memory.

While the details vary from language to language, we suppose here that we can choose variable names freely, subject only to the constraints that

- the first character must be a letter\(^2\) (a–z; A–Z),
- subsequent characters can be chosen from the letters (a–z; A–Z), the digits (0–9), and the underscore character (\_), and
- the several special words (keywords, e.g., IF, DO, UNTIL, ...) that we shall introduce in subsequent sections are reserved (or protected) and cannot be used as variable names.

While we do not explicitly impose a length limit on variable names, shorter names are preferred, simply because longer names take longer to type, take more space in the line, and, taken to extremes, result in program listings that are difficult to read. Generally, of course, we want to choose variable names that help us remember the significance of the quantity the name identifies.

\(^2\)Some languages are case-sensitive and others are not. For our general discussion, we shall assume case insensitivity, so A and a are indistinguishable in variable names.
Even in the generic discussion of this section, however, we shall pay attention to the data type of each quantity represented. In Chapter 1, we distinguished floating point numbers, integers, and character strings from one another in several ways. In structuring programs, we must remain always aware of this distinction. Further, the computer in interpreting or compiling our source code must either make assumptions or be told the data type of each variable we use. Some languages (e.g., Pascal, C) require an explicit declaration of variable names and associated data types before the name can be used. Other languages (e.g., BASIC, FORTRAN) exploit implicit data typing by looking to the composition of the variable name to determine the intended data type—though more recent versions of FORTRAN, for example, also admit explicit data typing. Still others (e.g., IDL, MATLAB, OCTAVE) assign a data type dynamically and automatically on the basis of the value assigned to the variable. In our general discussion, we shall adopt a convention that keeps us consciously aware of data type every time we use a variable name. Following the pattern actually used in BASIC, we shall take variable names with no appended suffix to represent floating point numbers, variable names with an appended percent sign \% to represent integers, and variable names with an appended dollar sign \$ to represent character strings. Occasionally, we shall use an appended at sign @ to stand for any one of these three possibilities (no suffix, \%, or \$). Thus, in the pseudocode we are defining, the names \textit{x}, \textit{position}, and \textit{field} identify memory cells for floating point values, the names \textit{I\%}, \textit{count\%}, and \textit{lower\ limit\%} are valid names for cells storing integers, the names \textit{word\$} and \textit{first\ name\$} are valid names for cells storing strings, and \textit{item@} identifies a quantity of any type.

The variables we have discussed so far are \textit{scalar} or \textit{unstructured} variables; each variable name stands for a single floating point number, integer, or character string. In scientific uses especially, we frequently want to refer to an aggregate of numbers (the three components of a position vector, the nine elements of a $3 \times 3$ matrix, ...). To that end, the \textit{array} is among the \textit{structured} variables available in all scientific programming languages. An array may be one-dimensional (a vector), two-dimensional (an $m \times n$ matrix), or even higher dimensional. As a structured entity, an array is identified by a single name, e.g., \textit{data}, \textit{values\%}, or \textit{names\$}, depending on the data type of the elements of the array. Individual elements are identified by attaching an integer index or indices to the array name, e.g., \textit{data(1\%, 3\%)}, \textit{values\%(4\%)}, or \textit{names\$(12\%)}. All by itself, the name of a \textit{scalar} is sufficient to tell the compiler how much memory to reserve to store the scalar value. For \textit{arrays}, however, we can't simply use the name when it is first needed. In most languages, the source code must also convey how large the array will become during the execution of the program because the interpreter or compiler must set aside adequate memory to store the array before the interpretation or compilation can be completed. Thus, a program that uses arrays must include in its source code a directive informing the compiler or interpreter of the maximum dimensions of the array. In our generic discussion, we shall use a statement like

\begin{verbatim}
DIMENSION data(4\%, 251\%), values\%(8\%), names\$(25\%)
\end{verbatim}

to convey the number of elements in each array to the interpreter or compiler. From the beginning, be aware that some standard languages by default use array indices that run upwards from the value 1\% while others start the indices at the value 0\%. In both cases, the statement above creates an array \textit{values\%}, for example, with eight elements. In the first case, the indices run from 1\% to 8\% while in the second case they run from 0\% to 7\%.

In C and in FORTRAN 90 and perhaps in a few other languages, it is possible to change the dimensions of an array dynamically, i.e., to bypass the need for the size of arrays to be explicitly specified in a \texttt{DIMENSION} statement and, instead, arrange for the size of an array to be set, for example, in response to values input when the program is executed. The coding to accomplish this \textit{dynamic allocation} of memory will be introduced if and when we find the need for it in the remainder of this book.
9.1.2 Essential Action Statements

The CPU must also be able to respond to a minimum collection of action statements to permit the assignment of values to variables, the performance of elementary arithmetic, and the transfer of information from and to a standard I/O device.\(^3\) We shall use the following special symbols and words for these operations:

- Assign a value to a named memory location:

\[
\text{(variable)} \leftarrow \text{(expression determining value to be assigned)}
\]

\(\text{e.g.,}\)\(^4\)

\[
\begin{align*}
\text{val} & \leftarrow \sin(\pi \times x) \\
\text{count}\% & \leftarrow \text{count}\% + 1\% \\
\text{first\_name}\$ & \leftarrow \text{"David"}
\end{align*}
\]

etc. Any valid arithmetic expression utilizing the symbols \(\text{+ for addition (or string concatenation), - for subtraction, \(* for multiplication, / for division, and }\text{^ for exponentiation can appear on the right hand side of the assignment symbol }\leftarrow.\) Further, most computer languages make available a wide variety of standard functions (\(\text{sin, cos, atan, sqrt, exp, }\ldots\)) to facilitate scientific computation.

- Obtain a value from the keyboard and store it in a (named) memory cell:

\[
\text{READ PROMPT=}\text{"Enter first name: }, \text{first\_name}\$
\]

- Transmit a value from a (named) memory cell and/or a (quoted) literal message to the screen:

\[
\text{WRITE }\text{"The value of the integral is }, \text{value}, \text{"."
}\]

We here recognize one further necessary refinement. In the above, we have simply requested information from the keyboard or directed output to the screen. Most programming languages will accept such simple statements, adopting a default \textit{format} in which to expect the input or produce the output. Almost always, we will wish to exert some control over that format, and the standard languages provide means to give us that control. At base, we want to be able to dictate how many character positions should be occupied by the number, whether the number is to be output as an integer or as a decimal number, whether the decimal number is to be presented in scientific or conventional notation, how many digits are to be placed after the decimal point, whether the number is to be rounded, \ldots Further, we may want to create blank lines in the output, ignore blank lines in the input, or clear the screen. Since this list is (almost) endless, the task of describing the format in which input will be presented to the computer and specifying the format in which the computer should produce its output is among the most complicated tasks the programmer confronts. Unfortunately, different languages adopt significantly different ways to provide this facility. Hence, detailed discussion beyond this brief recognition of a need is best postponed until we discuss specific languages.

\(^3\text{We confine our attention initially to data transfers from the keyboard and to the screen, postponing until later a discussion of the use of files.}\)

\(^4\text{We declare that integer }\textit{constants}\text{ shall be explicitly identified with a trailing percent sign and that string }\textit{constants}\text{ shall be enclosed in double quotation marks. Some languages use single quotation marks, while others will accept either (provided they are paired).}\)
9.1. COMPONENTS OF A PROGRAMMING LANGUAGE

Table 9.1: Simple logical expressions and the circumstances under which each is true. Remember that we have introduced the suffix @ to indicate any of the standard data types.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Truth if</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A@ = B@ )</td>
<td>( A@ ) is the same as ( B@ ) (( A@ ) equals ( B@ ))</td>
</tr>
<tr>
<td>( A@ &lt; B@ )</td>
<td>( A@ ) occurs before ( B@ ) in some collating sequence (e.g., numerical order, alphabetical order) (( A@ ) less than ( B@ ))</td>
</tr>
<tr>
<td>( A@ &gt; B@ )</td>
<td>( A@ ) occurs after ( B@ ) in that collating sequence (( A@ ) greater than ( B@ ))</td>
</tr>
<tr>
<td>( A@ &lt;&gt; B@ )</td>
<td>( A@ ) differs from ( B@ ) (( A@ ) not equal to ( B@ ))</td>
</tr>
<tr>
<td>( A@ &lt;= B@ )</td>
<td>( A@ ) occurs before or is the same as ( B@ ) (( A@ ) less than or equal to ( B@ ))</td>
</tr>
<tr>
<td>( A@ &gt;= B@ )</td>
<td>( A@ ) occurs after or is the same as ( B@ ) (( A@ ) greater than or equal to ( B@ ))</td>
</tr>
</tbody>
</table>

Table 9.2: Truth tables for OR and AND.

<table>
<thead>
<tr>
<th>Proposition</th>
<th>Truth Value</th>
<th>Proposition</th>
<th>Truth Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>true OR true</td>
<td>true</td>
<td>true AND true</td>
<td>true</td>
</tr>
<tr>
<td>true OR false</td>
<td>true</td>
<td>true AND false</td>
<td>false</td>
</tr>
<tr>
<td>false OR true</td>
<td>true</td>
<td>false AND true</td>
<td>false</td>
</tr>
<tr>
<td>false OR false</td>
<td>false</td>
<td>false AND false</td>
<td>false</td>
</tr>
</tbody>
</table>

9.1.3 Logical Expressions and Conditions

Frequently, the CPU will need to make a decision on the basis of currently available information, performing different tasks depending on that decision. Usually, these decisions are binary. At base, the computer decides whether a particular logical proposition is true or false. Further, the propositions so examined are usually cast as a comparison of two quantities to determine how they would be ordered in some standard ordering sequence (numeric, alphabetic, ...). The six simple comparisons we might want to code and the circumstances under which the computer will judge each to be true are enumerated in Table 9.1.

Two additional capabilities are standard in computer languages. First, we endow our CPU with the ability to interpret compound conditions constructed out of simpler conditions either with the operator OR or the operator AND. A truth table showing the result of connecting each possible logical value with each of these operators is presented in Table 9.2. Second, we introduce the ability to negate a condition with the operator NOT, i.e., we introduce the expression

\[ \text{NOT } \langle \text{condition} \rangle \]

which we define to be true if \( \langle \text{condition} \rangle \) is false and false if \( \langle \text{condition} \rangle \) is true.

Further (and finally), we recognize that we may occasionally find a need for a variable—a Boolean variable—that can assume only the two values true and false. Recognizing that different languages treat these variables differently, let us symbolize such a variable generically with a suffix ?. Thus, for example, we might at some point code a statement like

\[ \text{EQUAL?} \leftarrow \text{N@} = \text{M@} \]

in response to which the computer will assess whether \( \text{N@} \) is equal to \( \text{M@} \) and set \( \text{EQUAL?} \) to true or false, depending on the outcome of that assessment. With this expansion of our language, conditions might be expressed not only by the comparisons illustrated in Table 9.1 but also by the
simple assertion of a single Boolean variable or by a logical expression involving combinations of comparisons and/or assertions of Boolean variables. Note, incidentally, that the statement in this paragraph reveals why computer languages must have a different symbol for assignment than for testing equality.5

9.1.4 Syntactic Wrinkles

In some programming contexts, we will find it convenient—and sometimes even necessary—to convey additional information about the structure of the code to the CPU. We might wish

• to place more than one statement on a single physical line. In our generic code, we shall use the character ‘;’ to separate individual statements on a single line. The three lines of code early in Section 9.1.2 might be combined in one line with the coding

\[ \text{val } \leftarrow \sin(\pi \times x); \text{ count}\% \leftarrow \text{ count}\% + 1\%; \text{ first\_name$ } \leftarrow \text{ "David"} \]

• to spread a single statement over more than one physical line. In our generic code, we shall use the symbol → at the end of a line to indicate that the statement continues on the next line. Thus, for example, the two-line statement

\[ U(I\%, J\%) \leftarrow 0.25 \times ( U(I\%+1\%, J\%) + U(I\%-1\%, J\%) \rightarrow \\
+ U(I\%, J\%+1\%) + U(I\%, J\%-1\%) ) \]

is to be seen as a single statement.6

• to group several statements together to form a block of statements so that they can be properly treated when the block is placed in a context in which the interpreter or compiler is expecting a single statement. In our generic code, we shall use the keywords BEGIN BLOCK and END BLOCK to “bracket” the several statements that we wish the compiler or interpreter to see as a single (compound) statement. For example, the coding

\[ \begin{aligned}
\text{BEGIN BLOCK} \\
\text{val } \leftarrow \sin(\pi \times x) \\
\text{count}\% \leftarrow \text{ count}\% + 1\% \\
\text{first\_name$ } \leftarrow \text{ "David"} \\
\text{END BLOCK}
\end{aligned} \]

would group the three statements as a unit.

Different languages differ significantly in the way these three situations are coded. We shall be more explicit in subsequent sections as we address each language in turn.

9.1.5 Essential Control Structures

Four main control structures7 are generally provided in standard computer languages and figure significantly in the smooth expression of algorithms:

• Sequence, which is conveyed by the order of the statements in the algorithm.

---

5 We have used ← and ═ In later sections, we will identify the symbols used in other languages.

6 Remember that spaces not in quoted strings are ignored by the compiler.

7 Since the CASE structure can be expressed in terms of nested IF-THEN-ELSE structures and the IF-THEN-ELSE structure is simply a CASE structure with only two cases, there are really only three fundamental structures (sequence, selection, and repetition). General theorems in computer science prove that no task will require more than these few control structures.
• Two-way selection, expressed generically with a statement like\(^8\)\(^9\)

\[
\text{IF } \langle \text{condition} \rangle \\
\text{THEN } \langle \text{block1 of statements} \rangle \\
\text{ELSE } \langle \text{block2 of statements} \rangle \\
\text{END IF}
\]

Here \(\langle \text{block1 of statements} \rangle\) is executed when \(\langle \text{condition} \rangle\) is true and \(\langle \text{block2 of statements} \rangle\) is executed when \(\langle \text{condition} \rangle\) is false. The ELSE clause may be omitted altogether if \(\langle \text{block2 of statements} \rangle\) is null. The flow diagram in Fig. 9.1(a) is often used to convey this structure.

• Multi-way selection\(^10\)

\[
\text{CASE} \\
\text{OF } \langle \text{condition1} \rangle \text{ DO } \langle \text{block1 of statements} \rangle \\
\text{OF } \langle \text{condition2} \rangle \text{ DO } \langle \text{block2 of statements} \rangle \\
\text{OF } \langle \text{condition3} \rangle \text{ DO } \langle \text{block3 of statements} \rangle \\
\vdots \\
\text{OF OTHERS } \text{ DO } \langle \text{blockO of statements} \rangle \\
\text{END_CASE}
\]

In executing this overall statement, the computer will test each condition in turn, execute the block of statements associated with the first true condition it encounters, and then jump out of the CASE structure.\(^11\) The OTHERS clause may be omitted altogether if \(\langle \text{blockO of statements} \rangle\) is null. The flow diagram in Fig. 9.1(b) is often used to convey this structure.

• Repetition

\[
\text{LOOP} \\
\langle \text{block1 of statements} \rangle \\
\text{EXIT_LOOP WHEN } \langle \text{condition} \rangle \\
\langle \text{block2 of statements} \rangle \\
\text{END_LOOP}
\]

Here, the statements in the body of the loop are executed repeatedly until \(\langle \text{condition} \rangle\), which is tested at the indicated point in the loop, becomes true. Looping will continue forever unless repeated execution of the statements eventually causes \(\langle \text{condition} \rangle\) to become true. The flow diagram in Fig. 9.1(c) is often used to convey this structure.

In some more recent languages, the explicit construction of a loop is in some cases not necessary. For example, suppose we have a vector \(X\) containing \(N\) elements and we wish to create a second vector \(Y\), each of whose elements is—say—the sine of the corresponding element in \(X\). In the coding we have so far described, we would achieve this objective with an explicit loop involving the statements

\[
\text{CNT} \leftarrow 0 \quad ! \text{Set CNT to 0}
\]

for example, might as well be uncommented, since the comment really says no more than can be inferred from the statement itself.

\(^8\) The structure of this single statement is sufficiently distinctive—it is not complete until the keyword END_IF is encountered—that we need not use the symbol \(\rightarrow\) described in Section 9.1.4 to indicate that the statement is spread over several physical lines.

\(^9\) We use the exclamation point to introduce comments. Judicious use of such comments, irrelevant though they may be to the computer, can clarify the algorithm for the human reader of the listing. To be useful, a comment should amplify the meaning of, or clarify the role of, the commented statement. The statement

\[
\text{CNT} \leftarrow 0 \quad ! \text{Set CNT to 0}
\]

for example, might as well be uncommented, since the comment really says no more than can be inferred from the statement itself.

\(^10\) As with the previous structure, the structure of this single statement is sufficiently distinctive—it is not complete until the keyword END_CASE is encountered—that we need not use the symbol \(\rightarrow\) described in Section 9.1.4 to indicate that the statement is spread over several physical lines.

\(^11\) Not all compilers adopt this strategy. Some will execute the blocks associated with every true condition in the structure.
\begin{verbatim}
I% ← 0%
LOOP
  I% ← I% + 1%
  Y(I%) ← sin( X(I%) )
EXIT_LOOP WHEN I%=N%
END_LOOP
\end{verbatim}

(We assume that the vectors have already been appropriately dimensioned.) Were we working
with a language like IDL or MATLAB, which have built in array processing capabilities, this
loop would be automatically constructed in response to the single statement

\[ Y ← sin( X ) \]

These languages simply understand that, when functions like the sine are supplied with an
argument that is an array (whatever its dimension), the program is to compute an array of the
same dimension, each of whose elements is that function of the corresponding element in the
argument. These languages greatly simplify the coding of many operations involving arrays
and, in addition, almost certainly generate a more efficient execution of the entire task. When
available in the language in use, these capabilities should be exploited as much as possible.

Since blocks of statements may themselves involve any of these structures, the elementary structures
can give rise to more complicated structures in which elementary structures are nested two or more
deep. Note that, however complicated the THEN and ELSE clauses, the statements between CASE and
END_CASE, or the body of the loop, the (outer) IF-END_IF statement, the (outer) CASE-END_CASE
statement, and the loop as a whole are all seen by the compiler or interpreter logically as a single
statement.

Although the LOOP structure identified above is quite general and, by itself, is adequate to
accommodate all situations that might arise, structures that implement the terminating test in the
middle of the loop are rare in actual computer languages. Three alternative versions are usually
provided. In the first version, the test is at the beginning of the loop, e.g.,

\[ WHILE \langle\text{condition}\rangle \ DO \ (\text{block of statements}) \]

and the block is not executed at all if \langle\text{condition}\rangle is false when the loop is entered. In the second
version, the test is at the end of the loop, e.g.,

\[ REPEAT \ (\text{block of statements}) \ UNTIL \ (\text{condition}) \]

and the block will be executed at least once, whether \langle\text{condition}\rangle is true or false when the loop
is entered. Both of these versions will lead to infinite loops unless the statements in the body of
the loop ultimately toggle the condition to the value that will terminate the loop. Flow diagrams
depicting these structures are shown in Figs. 9.2(a) and (b).

The third version of a loop incorporates a built-in incrementation of an integer index and the
automatic testing of that index, e.g.,

\[ FOR \ I% ← IMIN% \ THRU \ IMAX% \ DO \ (\text{block of statements}) \]

When this loop is executed, the body of the loop is executed for \( I% \) having the value IMIN%, then
for \( I% \) having the value IMIN%+1%, then for \( I% \) having the value IMIN%+2%, and \ldots, continuing
until the loop has been executed for the largest value of \( I% \) that does not exceed IMAX%. The more
sophisticated form

\[ FOR \ I% ← IMIN% \ THRU \ IMAX% \ STEP \ INC% \ DO \ (\text{block of statements}) \]
9.1. COMPONENTS OF A PROGRAMMING LANGUAGE

Figure 9.1: Flow diagrams for the basic control structures: (a) two-way selection, (b) multi-way selection, (c) loop. Here, $T$, $F$, $B$, and $C$ abbreviate true, false, block, and condition, respectively.

![Flow diagrams for basic control structures](image)

Figure 9.2: Flow diagrams for different loop structures: (a) WHILE-DO, (b) REPEAT-UNTIL, (c) FOR-DO. Again, $T$, $F$, $B$, and $C$ abbreviate true, false, block, and condition, respectively.

![Flow diagrams for loop structures](image)

gives the user control over the increment by which the index is increased before each new pass through the body of the loop. A flow diagram for this loop structure is shown in Fig. 9.2(c). Note that the loop we have here depicted will be executed once if $\text{IMIN} \%$ equals $\text{IMAX} \%$ and not at all if $\text{IMIN} \%$ exceeds $\text{IMAX} \%$—but beware; different languages may behave differently in this regard.

Though it is not essential, one further control structure is standard in all scientific programming languages. Conveyed in our generic code by the statement

\[
\text{EXECUTE } \langle \text{procedure name} \rangle
\]
this single statement makes possible the invocation of a (properly defined) procedure in a program (or for that matter in another procedure); it both facilitates a modular approach to the design of programs and provides a means to avoid duplication of program modules that must be invoked more than once to specify a task completely.

9.2 Sample Short Algorithms

Large algorithms for accomplishing complex tasks are frequently constructed by combining various smaller algorithms. In this section, we enumerate several program fragments that may serve as building blocks for the construction of larger algorithms.

1. Exchange the values stored in A@ and B@:12

   `\begin{verbatim}
   TEMP@ ←− A@
   A@ ←− B@
   B@ ←− TEMP@
   \end{verbatim}`

   The temporary memory cell TEMP@ saves the original value of A@ so that it is still retrievable after the value of B@ has been copied into cell A@, overwriting the original contents of cell A@.

2. Stopping a loop with a sentinel [and counting]:

   `\begin{verbatim}
   SENTINEL@ ←− ⟨agreed-upon special value⟩
   [COUNT% ←− 0%]
   LOOP
      READ ITEM@
      EXIT_LOOP WHEN ITEM@ = SENTINEL@
         ⟨Statements processing ITEM@⟩
      [COUNT% ←− COUNT% + 1%]
   END_LOOP
   \end{verbatim}`

   If the statements enclosed in [...] are included, then when the loop is completed the variable COUNT% will have as its value the number of values of ITEM@ processed. Note that the position in the sequence at which COUNT% is incremented is critical. As a general rule, counters should be started at the value zero before anything happens and incremented by one immediately after each of the events to be counted. More often than not, the end result of thoughtless or unsystematic positioning of the incrementation will be a final count that is off by one, one way or the other.

3. Stopping a loop by counting up:

   `\begin{verbatim}
   NUMBER_OF_TIMES% ←− ⟨desired number of executions⟩
   COUNTER% ←− 0%
   LOOP
      EXIT_LOOP WHEN COUNTER% = NUMBER_OF_TIMES%
         ⟨block of statements⟩
      COUNTER% = COUNTER% + 1%
   END_LOOP
   \end{verbatim}

   Again, careful initialization of the counter and careful positioning of its incrementation are critical to avoiding off-by-one errors.

4. Stopping a loop by counting down:

   `[Note again the use of the character ! to flag comments. See footnote 9.]
TIMES_REMAINING% ← (desired number of executions)
LOOP
   EXIT_LOOP WHEN TIMES_REMAINING% = 0%
   (block of statements)
   TIMES_REMAINING% ← TIMES_REMAINING% - 1%
END_LOOP

This coding has a small advantage over the previous coding because it requires only one variable
(TIMES_REMAINING) to control the loop. Note, however, that the initial value of that variable
is irrecoverably lost by the time execution of the loop has been completed.

5. Summing [and counting]:

SENTINEL ← (agreed-upon special value)
SUM ← 0.0
[COUNT% ← 0%]
LOOP
   READ ITEM
   EXIT_LOOP WHEN ITEM = SENTINEL
   SUM ← SUM + ITEM
   [COUNT% ← COUNT% + 1%]
END_LOOP
WRITE "The sum is "; SUM

This algorithm for adding numbers involves the same steps you would use to accomplish the
task on a pocket calculator: initialize the accumulator to 0.0, enter each new value in turn,
push the ‘add’ button after each entry, stop after the last value has been processed, and
read the final value in the accumulator. Note the explicit decimal point in the floating point
constant 0.0. Since some compilers and interpreters in some circumstances will treat numerical
constants without explicit decimal points as integers, possibly producing unintended results,
prudence dictates habitually placing an explicit decimal point in all integer constants that
are in fact to be treated as floating point values. In the present situation, the variable SUM is
implicitly declared to be a floating point variable, so the statement SUM ← 0 would result in
an internal conversion of the integer value 0 to the floating point value 0.0, but it nonetheless
pays to be cautious.

6. Finding extreme, stopping with a sentinel:

SENTINEL@ ← (agreed-upon special value)
READ (first) ITEM@ from list
EXTREME@ ← ITEM@
LOOP
   READ (next) ITEM@ from list
   EXIT_LOOP WHEN ITEM@ = SENTINEL@
   IF ITEM@ and EXTREME@ are out of order
      THEN EXTREME@ ← ITEM@
   END_IF
END_LOOP
WRITE "The extreme value is "; EXTREME@

While this procedure can be executed with no a priori knowledge of the number of items and
no special assumptions about the list to come, it has the disadvantage of requiring the first
item of the list to be treated differently from the subsequent items. [Questions to the reader:
(1) Does this procedure behave sensibly if the sentinel is entered as the first item? How might

\footnote{For example, 5/2 = 2 in integer arithmetic; 5.0/2.0 = 2.5 in floating point arithmetic.}
7. Finding extreme, stopping by counting:

```
NUMBER_OF_VALUES% ← (number of items to be presented)
READ (first) ITEM@ from list
COUNTER% ← 1% ! Count item
EXTREME@ ← ITEM@ ! Assume first is extreme
LOOP
  EXIT_LOOP WHEN COUNTER% = NUMBER_OF_VALUES%
  READ (next) ITEM@ from list
  COUNTER% ← COUNTER% + 1%
  IF ITEM@ and EXTREME@ are out of order
    THEN EXTREME@ ← ITEM@
  END_IF
END_LOOP
WRITE "The extreme value is"; EXTREME@
```

[Question to the reader: What error would occur in the report if COUNTER% were incremented just before the EXIT_LOOP WHEN statement?]

8. Sequential search:

```
ITEM_Sought@ ← ⟨item to be found⟩
NO_ELEMENTS% ← ⟨number of items in ITEMS@⟩
POINT% ← 0% ! Start at beginning
LOOP
  POINT% ← POINT% + 1% ! Advance pointer to next item
  EXIT_LOOP WHEN ITEMS@(POINT%) = ITEM_Sought@ OR POINT%=NO_ELEMENTS%
END_LOOP
```

We assume that, by the time this procedure is invoked, the list of items to be examined has been stored in the one-dimensional array ITEMS and that we know the number of elements in that array. Note that the pointer POINT% is stepped through the records one at a time as the search unfolds. Finally, when this loop terminates, the item sought has been found if and only if the value in the record identified by POINT% matches the item sought.

### 9.3 Two Larger Algorithms

In this section, we discuss two algorithms in some detail, partly to illustrate the general features described in Section 9.1 more fully and partly to lay out the algorithms generically before we implement them in specific languages.

#### 9.3.1 Solving Laplace’s Equation

Among the more important equations in mathematical physics, Laplace’s equation appears in the study of electromagnetic fields, steady state heat flow, fluid mechanics, and many other contexts. In two-dimensions and in Cartesian coordinates \((x, y)\), the equation assumes the form

\[
\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = 0 \quad (9.1)
\]
for a function $U(x, y)$, which may be interpreted as an electrostatic potential, a temperature distribution, a velocity potential describing the incompressible, steady-state flow of a fluid, \ldots. Beyond the partial differential equation itself, a complete problem requires the statement of boundary conditions—often the stipulation of the value of $U$—at all points on the boundary of the region in which a solution is sought. Thus, for example, a complete problem might seek a solution to the Laplace equation in a square, subject to the requirement that the solution assume the value 0.0 on three edges of the square and the value 100.0 on the fourth edge, as shown in Fig. 9.3(a). Physically, this solution would convey the temperature within the square when three of its edges are maintained at 0° and the fourth edge is maintained at 100° or the electrostatic potential within the square when three edges are maintained at a potential of 0 Volts and the fourth edge is maintained at a potential of 100 Volts.

The basis for a simple algorithm for solving this problem numerically involves imposing an $N \times N$ regular grid of points $(x_i, y_j)$, $1 \leq i, j \leq N$, on the region. Then, we declare that we have found a solution when we know appropriate values for $U_{i,j} = U(x_i, y_j)$ at each grid point. The simplest boundary conditions, of course, tell us the values at all points for which $i$ and/or $j$ is either 1 or $N$. The values we seek for the other points ($2 \leq i, j \leq N - 1$) must in some sense reflect the differential equation. If, however, we know the values of $U$ at three consecutive points along a line parallel to the $x$ axis, say, we can use the approximation (see exercises)

$$\frac{\partial^2 U}{\partial x^2} \approx \frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{\Delta x^2}$$

where $\Delta x$ is the (constant) spacing of consecutive grid points. A similar expression applies as an approximation to the second derivative of $U$ with respect to $y$. Discretizing Eq. (9.1) by substituting these finite-difference approximations for the derivatives and then rearranging the resulting equation, we conclude that

$$U_{i,j} \approx \frac{1}{4} \left( U_{i+1,j} + U_{i-1,j} + U_{i,j+1} + U_{i,j-1} \right)$$

i.e., that the value we should assign to the “squared” point in Fig. 9.3(b) is the average of the values we assign to its four nearest neighbors (the “circled” points).

We, of course, have one equation like Eq. (9.3) for each of the interior points in the illustrated grid. We also have exactly as many unknowns as we have interior points. Further, the equations are linear. Thus, we have deduced a (probably large) set of linear equations to be solved simultaneously for the unknown values $U_{i,j}$ at the interior points. For this set of equations, a suitable strategy—called relaxation—­involves guessing a starting solution and then refining that solution iteratively by stepping systematically and repeatedly through the grid of interior points, replacing the value at
each point by the average of the values at its four nearest neighbors. Each pass through the entire grid constitutes one iteration, and we keep going—iteration after iteration—until we are satisfied that the process has converged satisfactorily.\footnote{The most common criterion of convergence involves comparing values until no element in one iterate differs from its counterpart in the next iterate by more than some predetermined tolerance. Developing suitable (and reliable) criteria for determining when satisfactory convergence has been achieved, however, is difficult. Because we are here focusing on programming aspects rather than algorithmic refinements, we shall ignore the issue of convergence for the moment. We shall return to it at appropriate points in later chapters.} Supposing that we seek a solution on a $15 \times 15$ grid of points, we make a first pass at constructing an algorithm to implement this overall strategy by listing the statements

```plaintext
DIMENSION U(15%,15%)
Set U(I%,J%) ←− 0.0 for all values of I%, J%
Set U(15%,J%) ←− 100.0 for all values of J%
FOR ITCNT% ←− 1% THRU 30% DO Conduct one iteration
Write solution to output device
```

Here, we (1) reserve adequate space for the necessary array, (2) establish a starting solution in that array, (3) set the boundary conditions, (4) conduct 30 iterations, and (5) display the solution.

To refine the crude statements in the above algorithm, we would have to expand the implied loops. Indeed, since we are working with a two-dimensional array, stepping through all elements of the array (either in assigning initial values or in iterating once through the array) will require a double loop. Thus, we might express this algorithm more explicitly as

```plaintext
DIMENSION U(15%,15%)
FOR I% ←− 1% THRU 15% DO
    FOR J% ←− 1% THRU 15% DO U(I%,J%) ←− 0.0
FOR I% ←− 1% THRU 15% DO
    U(15%,J%) ←− 100.0
    ! Prepare array of proper size
    ! Set all elements of array to 0.0
    ! Correct values on right edge
FOR ITCNT% ←− 1% THRU 30% DO
    ! Iterate 30 times
    FOR I% ←− 2% THRU 14% DO
        FOR J% ←− 2% THRU 14% DO
            U(I%,J%) ←− 0.25 * ( U(I%+1%, J% →
                               + U(I%-1%, J%) + U(I%, J%+1%) →
                               + U(I%, J%-1%) )
        FOR I% ←− 1% THRU 15% DO
            FOR J% ←− 1% THRU 15% DO WRITE U(I%,J%)
    ! Conduct one iteration
    ! Write solution to output device
```

We shall refine this algorithm even further as, later in this chapter, we implement it in various computer languages.

### 9.3.2 File Output/Input

The most convenient and flexible way to make a permanent record of numerical data is to store it in an ASCII file. Among other advantages, such files can be created in any number of ways, they can be examined with ordinary text editors, and they facilitate importing the data into whatever graphical package provides the best visualization of the specific features we wish to see. Most often the data are naturally organized into one or more two- or three-dimensional arrays, which frequently represent scalar or vector fields. In this section, we declare a standardized file format for the storage of these arrays and discuss a general algorithm by which files so structured can be created. In later sections, we describe how to implement this algorithm in various computer languages. Of course, the data file as an intermediary is unnecessary if generation and display occur in the same program. These files are especially useful when data generated in one program are to be transferred to another program for display and/or when a program-independent permanent record of the data is desired.
9.3. TWO LARGER ALGORITHMS

Table 9.3: Sample data file having the structure described in Section 9.3.2. The line numbers in the first column and the comments in the third column of this table are not in the actual file.

<table>
<thead>
<tr>
<th>Line Number</th>
<th>Line in file</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>01:</td>
<td>Example Data File; Author: DMC; Date: 7-21-01</td>
<td>ID of file</td>
</tr>
<tr>
<td>02:</td>
<td>First line of comments.</td>
<td></td>
</tr>
<tr>
<td>03:</td>
<td>Second line of comments.</td>
<td></td>
</tr>
<tr>
<td>04:</td>
<td>Third line of comments.</td>
<td></td>
</tr>
<tr>
<td>05:</td>
<td>Fourth line of comments.</td>
<td></td>
</tr>
<tr>
<td>06:</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>07:</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>08:</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>09:</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>10:</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>11:</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>12:</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>13:</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>14:</td>
<td>0.66</td>
<td></td>
</tr>
<tr>
<td>15:</td>
<td>0.57</td>
<td></td>
</tr>
</tbody>
</table>

At base, we imagine that the data files we create will store some number $n$ of identically dimensioned three-dimensional arrays $A_{ijk}$ whose three indices will assume $n_x$, $n_y$, and $n_z$ values,\(^{15}\) respectively, though at times $n_z$ will have the value 1 so that the storage of two-dimensional arrays can be accommodated. We adopt the following standard format for all data files:

- five lines of comments, possibly including title, author, and date;
- one line specifying the number $n$ of arrays in the file;
- one line specifying the number $n_x$ of values assumed by the first index $i$ of $A_{ijk}$;
- one line specifying the number $n_y$ of values assumed by the second index $j$ of $A_{ijk}$;
- one line specifying the number $n_z$ of values assumed by the third index $k$ of $A_{ijk}$; and finally
- $n_x n_y n_z$ lines specifying the values in the first array—one value per line in the order that results from the generic coding

\[
\text{FOR } K\% \leftarrow 1\% \text{ THRU } NZ\% \text{ DO}
\]
\[
\text{FOR } J\% \leftarrow 1\% \text{ THRU } NY\% \text{ DO}
\]
\[
\text{FOR } I\% \leftarrow 1\% \text{ THRU } NX\% \text{ DO WRITE } A(I\%, J\%, K\%)
\]

—followed by a similarly structured presentation of the values in the second array, the third array, etc. That is, the order of elements for each array is created by allowing the first index to vary the most rapidly, the second index to vary next rapidly, and the third index to vary least rapidly.

The data file shown in Table 9.3 is a simple example. This file contains one $2 \times 3 \times 1$ (2 rows $\times$ 3 columns $\times$ 1 plane) array, namely

\[
A = \begin{pmatrix}
A_{111} & A_{121} & A_{131} \\
A_{211} & A_{221} & A_{231}
\end{pmatrix} = \begin{pmatrix}
0.33 & 0.50 & 0.66 \\
0.28 & 0.42 & 0.57
\end{pmatrix}
\]

with the elements in the first column, then the elements in the second column, and finally the elements in the third column recorded in the file.

\(^{15}\)In some languages, the actual values of these indices will range from 0 to $n_\ast - 1$; in other languages, the values will range from 1 to $n_\ast$. (The subscript $\ast$ stands for $x$, $y$, or $z$.) In a few languages, the user has control over the range of the indices.
To be even more explicit, we would have to surround the basic output statement presented above with statements that dimension arrays appropriately, calculate (or otherwise determine) the values to insert in the arrays, prepare the file for access, and close the file after the last datum has been written to the file. To illustrate, suppose we seek to explore the two-dimensional scalar field given in the $xy$ plane by the function $f(x, y)$. Then, generating the desired file involves two main steps: (1) creation of an internal array containing values of the function at a grid of points overlayed on the region of interest in the $xy$ plane and (2) writing those values into a suitably labeled file structured as described earlier in this section. Though it may take a bit of exploration to decide on suitable ranges for the independent variable, let us here decide to examine the function in the region defined by $x_{\text{min}} \leq x \leq x_{\text{max}}$ and $y_{\text{min}} \leq y \leq y_{\text{max}}$ and to divide the $x$ interval with $n_x$ grid points into $n_x - 1$ segments and the $y$ interval with $n_y$ grid points into $n_y - 1$ segments. First, we prepare variables and arrays with the statements

\[
\begin{align*}
\text{DIMENSION ARRAY}(n_x, n_y) & \quad ! \text{Prepare array for values} \\
\text{NARR}\% & \leftarrow 1\% \quad ! \text{Set number of arrays} \\
\text{NX}\% & \leftarrow n_x \quad ! \text{Set number of grid points} \\
\text{NY}\% & \leftarrow n_y \quad ! \text{in each coordinate} \\
\text{NZ}\% & \leftarrow 1\% \quad ! \text{Set increments} \\
\text{DX} & \leftarrow (x_{\text{max}} - x_{\text{min}})/(n_x - 1) \\
\text{DY} & \leftarrow (y_{\text{max}} - y_{\text{min}})/(n_y - 1)
\end{align*}
\]

Then, we evaluate the function $f(x, y)$ with the double loop expressed in the code

\[
\begin{align*}
\text{FOR J}\% & \leftarrow 1\% \text{ THRU NY}\% \text{ DO} \\
\text{BEGIN BLOCK} & \\
\text{YF} & \leftarrow y_{\text{min}} + (J\%-1\%)*\text{DY} \\
\text{FOR I}\% & \leftarrow 1\% \text{ THRU NX}\% \text{ DO} \\
\text{BEGIN_BLOCK} & \\
\text{XF} & \leftarrow x_{\text{min}} + (I\%-1\%)*\text{DX} \\
\text{ARRAY}(I\%, J\%) & \leftarrow f(\text{XF}, \text{YF}) \\
\text{END BLOCK} & \\
\text{END BLOCK} & \\
\end{align*}
\]

Finally, we establish communication between the file and the program (in the jargon, we attach the file to the program on a selected channel), write the necessary labels and values to the file, and disconnect (i.e., detach the file from the program with the statements\(^{16}\)

\[
\begin{align*}
\text{ATTACH FILE (filename) ON CHANNEL 1 FOR WRITING} & \\
\text{WRITE TO CHANNEL 1, "Line of explanation (Title, Author, Date?)"} & \\
\text{WRITE TO CHANNEL 1, "Line of explanation"} & \\
\text{WRITE TO CHANNEL 1, "***"} & \\
\text{WRITE TO CHANNEL 1, "***"} & \\
\text{WRITE TO CHANNEL 1, NARR\%} & \\
\text{WRITE TO CHANNEL 1, NX\%} & \\
\text{WRITE TO CHANNEL 1, NY\%} & \\
\text{WRITE TO CHANNEL 1, NZ\%} & \\
\text{FOR J\% \leftarrow 1\% TO NY\% DO FOR I\% \leftarrow 1\% TO NX\% DO WRITE TO CHANNEL 1, ARRAY(I\%, J\%)} & \\
\text{CLOSE FILE ON CHANNEL 1} & \\
\end{align*}
\]

Here, the first statement identifies the file and assigns to it a number—any number will do\(^{17}\)—to be used for subsequent reference to the file. Next, the first five WRITE statements send five lines of

\(^{16}\)A new file will be created, overwriting any existing file by the specified name. Some programming languages may provide a warning if an existing file will be deleted. 

\(^{17}\)Be aware that some languages reserve a few channel numbers for special "files", e.g., the keyboard or the screen.
9.10 SOLVING LAPLACE’S EQUATION WITH OCTAVE

Comments, all five of which must be physically present even if fewer than five are needed to contain necessary information about the file. Then, the sixth WRITE statement writes the number of arrays to the file, the seventh and eighth write the x and y dimensions to the file, and the ninth writes the z dimension to the file. Finally, the double loop writes the elements of the array to the file in the proper order, and the last statement makes sure the file is properly detached from the program so that it can be accessed by other programs.

Instead of writing data to a file, we frequently will need to read data from a file, perhaps as a means to import data generated by one program for use in a different program. Slight modification of the ATTACH and READ statements introduced earlier provide for that action. Specifically, the statements

ATTACH FILE ⟨filename⟩ ON CHANNEL 1 FOR READING
READ FROM CHANNEL 1, Comma-separated list of variables for storage of values

; 
CLOSE FILE ON CHANNEL 1

Here, the first statement opens an existing file for reading (and displays an error if the specified file does not exist), the remaining statements save the last read the information in the file and stores it in the specified variables, and the last statement detaches the file when, presumably all data have been read. Note that the structure of the READ statements must reflect the structure of the file, and the variables specified must have the data types of the values to be read. Normally, the READ statement reads a line at a time, but in some languages each READ statement reads as many values as needed to complete the list of variables, however these values are distributed over lines.

In the remaining sections of this chapter, we shall implement the two algorithms described in this section in different languages and, especially for the second algorithm, in a variety of different contexts.

9.10 Solving Laplace’s Equation with OCTAVE

The algorithm described in Section 9.3.1 is readily implemented in OCTAVE with the statements

```
>> u = zeros(15,15); u(:,15)=100.0;
>> for itcnt = 1:30
    for i = 2:14
      for j = 2:14
        u(i,j) = 0.25* ( u(i+1,j)+u(i-1,j)+u(i,j+1)+u(i,j-1) );
      end
    end
  end
```

Here, the two statements on the first line establish a $15 \times 15$ matrix of zeros and adjust the values in the final column to the stipulated boundary values. Then, the remaining statement (which contains three nested for loops and occupies seven physical lines) executes 30 iterations (the outermost for loop) of the double loop it contains. The innermost statement simply expresses the substance of Eq. (9.3). Note that the inner loops do not include points on the boundary of the region in which a solution is sought, so the values in the array u at its “edges” are left alone throughout the 30 iterations.

Once the solution has been completed, of course, we could display it on the screen with any number of statements, the simplest of which probably are
>> mesh( u )

or

>> surf( u )

though these could be embellished by exploiting additional arguments and commands to rescale the axes, select alternate colors, label the display, etc.

9.12 Creating and Storing Two-Dimensional Scalar Arrays

We turn now to the storage of one or more arrays representing scalar and vector fields in files with the structure described in Section 9.3.2. Suppose, first, that we wish to explore the two-dimensional scalar field representing the irradiance produced by Fraunhofer diffraction at a square aperture and given analytically by the expression

\[ I(x,y) = I_0 \left( \frac{\sin x}{x} \right)^2 \left( \frac{\sin y}{y} \right)^2 \]  

(9.4)

Here, \( I_0 \) is the irradiance at the center of the pattern \((x = y = 0)\). Let us here decide that we seek values of \( I(x,y)/I_0 \) over the region \(-3\pi \leq x, y \leq 3\pi\), and that we shall divide each axis into 49 segments of length \( 6\pi/49 \), which will entail evaluating \( I(x,y)/I_0 \) at 50 values of \( x \) and 50 values of \( y \)—a total of \( 50 \times 50 = 2500 \) values.\(^{18}\)

9.12.5 ... with OCTAVE

\(^{18}\)Note that the function to be evaluated is ill-defined at \( x = 0 \) and at \( y = 0 \). We choose 49 rather than 50 segments so as to avoid having some of the points at which the function is evaluated fall at \( x = 0 \) or \( y = 0 \). Were the function well defined everywhere, such an awkward number of divisions would not be necessary.

\(^{19}\)Remember that, in OCTAVE, array indices start at 1, so division of the interval into 49 segments will result in 50 grid points along each axis and in indices running from 1 to 50.

\(^{20}\)A commented listing of this m-file is presented in Section 9.G.

\[ \text{inc} = 6.0 \cdot \pi/49.0; \]

\[ \text{[x,y]} = \text{meshgrid}(-3.0 \cdot \pi : \text{inc}:3.0 \cdot \pi, -3.0 \cdot \pi : \text{inc}:3.0 \cdot \pi); \]

\[ \text{I} = ( (\sin(x) ./x)^{\cdot 2} ) \cdot ( (\sin(y) ./y)^{\cdot 2} ); \]

This array can be written to a file having the format described in Section 9.3.2 by using the Lawrence-written M-file \texttt{lugen_list.m}.\(^{20}\) The statement

\[ \text{lugen_list(I)} \]

calls \texttt{lugen_list} and tells OCTAVE to write the array \texttt{I}, which must already exist in the OCTAVE workspace, to a file. Before accomplishing its main objective, \texttt{lugen_list} requests input of

\[ \text{Note: All OCTAVE programs (*.m) and all OCTAVE-created data files (*.octave.dat) in this chapter can be copied from the directory $\texttt{HEAD/octave}$, where (as defined in the Local Guide) $\texttt{HEAD}$ must be replaced by the appropriate path for your site.} \]
• the name of the file into which the array is to be written,
• a brief title for the data,
• the author’s (owner’s) name,
• the date, and
• four lines of comments.

Each of these pieces of information is a string, which therefore must be enclosed in single quotation marks as it is entered in response to its prompt. All default strings that are not null are indicated in square brackets in the prompting messages. For example, the “conversation” that would save the values of irradiance in the array I might be

```matlab
>> lugen_list(I)
```

All entries are strings and must be enclosed in single quotation marks.

File name [untitled.dat]: 'irrad_octave.dat'
Title : 'Irradiance'
Author : 'David M. Cook'
Date : '19 June 2001'
Comment 1 [**]: 'Fraunhofer diffraction at square aperture'
Comment 2 [**]: 'Program described in CPSUP'
Comment 3 [**]:
Comment 4 [**]:

File 'irrad_octave.dat' saved successfully!

When ⟨RETURN⟩ is typed after the fourth line of comments, `lugen_list` writes the data into the specified file in the default directory and displays a message indicating successful completion of that action on the screen.\(^{21}\)

### 9.13 Creating and Storing Three-Dimensional Scalar Arrays

As an example of a three-dimensional scalar function, we choose the normalized probability density \(p(x, y, z)\) for the electron in the \((n, l, m) = (3, 2, 0)\) state in the hydrogen atom. This probability density is given as a function of Cartesian coordinates \((x, y, z)\) with the nucleus located at the origin by

\[
p(x, y, z) = \frac{1}{2\pi(27)^3} e^{-2\rho/3} \left( 3 \frac{z^2}{\rho^2} - 1 \right)^2 = \frac{1}{2\pi(27)^3} e^{-2\rho/3} (9z^4 - 6z^2\rho^2 + \rho^4) \tag{9.5}
\]

where the coordinates are all measured in units of the Bohr radius and

\[
\rho = \sqrt{x^2 + y^2 + z^2} \tag{9.6}
\]

This function of three variables is commonly visualized either by focusing on the function in various planes intersecting the three-dimensional volume (thereby reducing the display to a family of two-dimensional displays) or by displaying various contour surfaces. Programs producing such displays need a three-dimensional array of values as input. Suppose, then, we anticipated using some graphical visualization program to explore the quantum probability given by Eq. (9.6). Let us decide to determine values of \(p(x, y, z)\) over the region \(-10 \leq x, y, z \leq 10\), dividing each axis into 29 segments,

\(^{21}\) A Lawrence-written OCTAVE procedure for reading files written by `lugen_list.m` is described in Section 9.16.5.
which will entail evaluating \( p(x, y, z) \) at 30 values of \( x \), 30 values of \( y \), and 30 values of \( z \)—a total of 
\( 30 \times 30 \times 30 = 27000 \) values.

Since the programs discussed in this section differ very little from those presented in Section 9.12, we elect here to include full code only in the listings in the appendices. We comment in the text only on major differences and subtleties.

9.13.5 ... with OCTAVE

The procedure for using OCTAVE to generate three-dimensional scalar arrays is similar to the method described in Section 9.12.5. The array representing the probability density in Eq. (9.5) is created using the OCTAVE statements

\[
\begin{align*}
&\text{inc} = 20.0/29.0; \\
&[x,y,z] = \text{meshgrid}(-10.0:\text{inc}:10.0, -10.0:\text{inc}:10.0, -10.0:\text{inc}:10.0); \\
&p\text{factor} = 1.0/(2.0*\pi*27.0*27.0*27.0); \\
&zs = z.*z; \\
&rhos = x.^2 + y.^2 + zs; \\
&rho = \text{sqrt}(\text{rho}); \\
&P = p\text{factor} \times \exp(-2*\text{rho}/3) .\times (9*zs.^2 - 6*zs.*\text{rho} + \text{rho}^2);
\end{align*}
\]

The first two lines produce a grid of values for the evaluation of the scalar field. Here, \( x \), \( y \), and \( z \) each ranges from \(-10.0\) to \(10.0\), and the interval is divided into 29 segments, so the arrays \( x \), \( y \), \( z \), \( zs \), \( rhos \), \( rho \), and \( P \) have dimensions \( 30 \times 30 \times 30 \). The remaining lines evaluate several intermediate quantities and then evaluate \( P \). Finally, with the statement

\[
\text{lugen\_list}\left(P\right)
\]

we write the data into a file in the desired format.\(^{22}\) After the filename—we here choose \texttt{pdens\_octave.dat}—and other labeling information has been typed in response to the sequence of prompts, the data file will be saved in the default directory.\(^{23}\)

9.14 Creating and Storing Two-Dimensional Vector Arrays

Display of a vector field in two dimensions requires a two-dimensional vector array, i.e., a two-dimensional array, each of whose elements is itself a vector. Instead of creating a single two-dimensional vector array, however, we elect to construct a pair of two-dimensional scalar arrays, one for each component of the vector. The first array contains the first component of the vector field at each point on a grid covering the region of interest and the second array contains the second component of the vector field on that same grid.

As an example of a two-dimensional vector field, we choose the magnetic field in a transverse electric (TE) electromagnetic wave propagating in a waveguide with perfectly conducting walls (though we choose also to compute and save the associated electric field). As shown in Fig. 9.4, we take the waveguide to be oriented so that the wave propagates in the positive \( x \) direction. Let the guide have a rectangular cross-section with dimensions \((b, d)\), i.e., \((0 < y < b, 0 < z < d)\). Further,
let the wave be polarized in the $z$ direction and consider the particular TE waves whose electric field does not depend on $z$. In mksa units, the (complex) fields in this guide are given by\(^{24}\)

\[
\mathbf{E}(\mathbf{r}, t) = A \sin \frac{n\pi y}{b} e^{i(\kappa_x x - \omega t)} \hat{k} \tag{9.7}
\]

\[
\mathbf{H}(\mathbf{r}, t) = \frac{\nabla \times \mathbf{E}}{i\omega \mu_0} = \frac{A}{i\omega \mu_0} \left( \frac{n\pi}{b} \cos \frac{n\pi y}{b} \hat{i} - i\kappa_x \sin \frac{n\pi y}{b} \hat{j} \right) e^{i(\kappa_x x - \omega t)} \tag{9.8}
\]

where $n = 1, 2, 3, \ldots$ and, with $c$ standing for the speed of light,

\[
\kappa_x^2 = \frac{\omega^2}{c^2} - \left( \frac{n\pi}{b} \right)^2 \tag{9.9}
\]

If we focus on the physical fields (real parts of $\mathbf{E}$ and $\mathbf{H}$) specifically at $t = 0$, we find that

\[
\mathbf{E}(\mathbf{r}) = A \cos \kappa_x x \sin \frac{n\pi y}{b} \hat{k} \tag{9.10}
\]

\[
\mathbf{H}(\mathbf{r}) = \frac{A n\pi}{\omega \mu_0 b} \sin \kappa_x x \cos \frac{n\pi y}{b} \hat{i} - \frac{A \kappa_x}{\omega \mu_0} \cos \kappa_x x \sin \frac{n\pi y}{b} \hat{j}
\]

To simplify these equations further, we divide $\mathbf{E}$ by $A$ and $\mathbf{H}$ by $A n\pi/\omega \mu_0 b$, and we elect to measure lengths in units of $b$ by introducing the variables $\overline{y} = y/b$ and $\overline{x} = x/b$. Equation (9.10) then becomes

\[
\overline{\mathbf{E}}(\mathbf{r}) = \frac{\mathbf{E}(\mathbf{r})}{A} = \cos \kappa_x b \overline{x} \sin n\pi \overline{y} \hat{k} \tag{9.11}
\]

\[
\overline{\mathbf{H}}(\mathbf{r}) = \frac{\mathbf{H}(\mathbf{r})}{A n\pi/\omega \mu_0 b} = \sin \kappa_x b \overline{x} \cos n\pi \overline{y} \hat{i} - \frac{\kappa_x b}{n\pi} \cos \kappa_x b \overline{x} \sin n\pi \overline{y} \hat{j}
\]

and we find that the field of interest depends on two parameters $n$ and $\kappa_x b$. Further, Eq. (9.9) imposes the constraint

\[
\left( \frac{\kappa_x b}{n\pi} \right)^2 = \left( \frac{\omega b}{n\pi c} \right)^2 - 1 \tag{9.12}
\]

\(^{24}\)For a derivation of the magnetic field for a wave propagating in the $z$ direction and a discussion on wave guides in general, see *The Theory of the Electromagnetic Field* by David M. Cook (Prentice-Hall, Englewood Cliffs, NJ, 1975) or *Introduction to Electrodynamics* by David J. Griffiths (Prentice-Hall, Upper Saddle River, NJ, 1999), Third Edition. The first of these books, out of print since the early 1990’s, became available in a Dover reprint in January, 2003, but is now (January 2017) also out of print. It may still be available in your local library.
on these parameters. The only non-zero components of the fields in Eq. (9.11) are

\[
\begin{align*}
\mathbf{E}_z &= \cos \kappa_x b \frac{x}{b} \cos n \pi \frac{y}{b} \\
\mathbf{H}_x &= \sin \kappa_x b \frac{x}{b} \cos n \pi \frac{y}{b} \\
\mathbf{H}_y &= -\frac{\kappa_x b}{n \pi} \cos \kappa_x b \frac{x}{b} \cos n \pi \frac{y}{b}
\end{align*}
\] (9.13)

For the sake of a specific example, we choose \( b = 1 \) and \( n = 2 \), and then we choose \( \omega \) so that \( \kappa_x b / 2\pi \) turns out to have the value 1. With these choices, we find that the fields we seek to display are given by

\[
\begin{align*}
\mathbf{E}_z &= \cos 2 \pi \frac{x}{b} \sin 2 \pi \frac{y}{b} \\
\mathbf{H}_x &= \sin 2 \pi \frac{x}{b} \cos 2 \pi \frac{y}{b} \\
\mathbf{H}_y &= -\cos 2 \pi \frac{x}{b} \sin 2 \pi \frac{y}{b}
\end{align*}
\] (9.14)

where \( x \) can range over any values—we choose \( 0 \leq x \leq 1 \)—but, to be inside the guide, \( y \) is confined to the region \( 0 \leq y \leq 1 \). Each component of these fields can now be represented by a two-dimensional array. The \( \mathbf{H} \) field, which has two non-zero components, then is translated into two such arrays; the \( \mathbf{E} \) field, which has only one non-zero component, requires only one such array. To be explicit, we determine values of \( \mathbf{H}(x,y) \), \( \mathbf{E}(x,y) \), and \( \mathbf{E}(x,y) \) over the region \( 0 \leq x, y \leq 1 \), dividing each axis into 30 segments of length \( 1.0 / 29.0 \), which will entail evaluating the field components at 30 values of \( x \) and 30 values of \( y \)—a total of \( 30 \times 30 = 900 \) values. Normally, these field components would be stored in three \( 30 \times 30 \) two-dimensional arrays. Because of our declared file format, however, we must view ourselves as needing storage for three \( 30 \times 30 \times 1 \) three-dimensional arrays. We shall, however, view the structure as a single \( 30 \times 30 \times 1 \times 3 \) array \( \mathbf{H} \), with \( \mathbf{H}(*,*,1,1) \) storing \( \mathbf{H}_x \), \( \mathbf{H}(*,*,1,2) \) storing \( \mathbf{H}_y \), and \( \mathbf{H}(*,*,1,3) \) storing \( \mathbf{E}_z \).

9.14.5 ... with OCTAVE

The procedure for generating two-dimensional vector arrays is very similar to the method described in Section 9.12.5. However, in this case, we have three two-dimensional arrays, one for each component of the vector fields. The arrays representing the fields in Eq. (9.14) are created using the OCTAVE commands

\[
\begin{align*}
\text{>> inc} &= 2.0 \times \text{pi} / 29.0; \\
\text{>> [x,y]} &= \text{meshgrid( 0.0:inc:2.0*pi, 0.0:inc:2.0*pi )}; \\
\text{>> H}(::,1,1) &= \sin(x) \times \cos(y); \\
\text{>> H}(::,1,2) &= -\cos(x) \times \sin(y); \\
\text{>> H}(::,1,3) &= -\text{H}(::,1,2);
\end{align*}
\]

The first two lines produce a grid of values for the evaluation of each component of the fields. Here, \( x \) ranges from 0 to \( 2\pi \), \( y \) ranges from 0 to \( 2\pi \), and the interval is divided into 29 segments. The third, fourth, and fifth lines evaluate the two components of \( \mathbf{H} \) and the one component of \( \mathbf{E} \), stored respectively in \( \mathbf{H}(::,1,1), \mathbf{H}(::,1,2), \) and \( \mathbf{H}(::,1,3) \). Finally, with the statement

\[
\begin{align*}
\text{>> lugen_list( H )}
\end{align*}
\]
9.15. CREATING AND STORING THREE-DIMENSIONAL VECTOR ARRAYS

we write the data into a file in the desired format. After the filename—we here choose \texttt{wavedg.octave.dat}—and other labeling information has been typed in response to the sequence of prompts, the data file will be saved in the default directory.

9.15 Creating and Storing Three-Dimensional Vector Arrays

Display of a vector field in three dimensions requires a three-dimensional vector array, i.e., a three-dimensional array, each of whose elements is itself a vector. Instead of creating a single three-dimensional vector array, however, we elect to construct a triplet of three-dimensional scalar arrays, one for each component of the vector. The first array contains the first component of the vector field at each point on a grid covering the region of interest, the second array contains the second component of the vector field on that same grid, and the third array contains the third component of the vector field. This section describes convenient ways to produce such triplets of arrays and to write them into ASCII files for transfer to other programs.

As an example of a three-dimensional vector field, we choose the electric field produced by a quadrupole consisting of four charges at the corners of a square of side $2a$ with its center at the origin and its plane in the $xy$ plane. Choosing to measure the coordinates $x$, $y$ and $z$ in units of $a$, we find that this field is given by

$$\frac{\left| E_x(x, y, z), E_y(x, y, z), E_z(x, y, z) \right|}{q/4\pi\varepsilon_0 a^2} = \left[ \begin{array}{c} \frac{[x-1,y-1,z]}{[(x-1)^2+(y-1)^2+z^2]^{3/2}} - \frac{[x+1,y-1,z]}{[(x+1)^2+(y-1)^2+z^2]^{3/2}} \\ \frac{[x+1,y+1,z]}{[(x+1)^2+(y+1)^2+z^2]^{3/2}} - \frac{[x-1,y+1,z]}{[(x-1)^2+(y+1)^2+z^2]^{3/2}} \\ \frac{[x+1,y+1,z]}{[(x+1)^2+(y+1)^2+z^2]^{3/2}} - \frac{[x-1,y+1,z]}{[(x-1)^2+(y+1)^2+z^2]^{3/2}} \end{array} \right]$$

(9.15)

To be explicit, we determine values of the field components over the region $-2.0 \leq x, y, z \leq 2.0$, dividing each axis into 29 segments, which will entail evaluating the components at 30 values of $x$, 30 values of $y$, and 30 values of $z$—a total of $30 \times 30 \times 30 = 27000$ values. Because of our declared file format, however, we must view ourselves as needing storage for three $30 \times 30 \times 30$ three-dimensional arrays. We shall, however, view the structure as a single $30 \times 30 \times 30$ array $E$, with $E(*)**,*,1)$ storing $E_x$, $E(*,**,**,2)$ storing $E_y$, and $E(*,**,**,3)$ storing $E_z$.

Since the programs discussed in this section differ very little from those presented in Section 9.14, we elect here to include full code only in the listings in the appendices. We comment in the text only on major differences and subtleties.

9.15.5 \ldots with OCTAVE

The procedure for using OCTAVE to generate three-dimensional vector arrays is very similar to the method described in Section 9.14.5. In this case, however, we must create three three-dimensional arrays, corresponding to the components of the three-dimensional vector. The array representing the components of the $E$-field given by Eq (9.15) is created using the OCTAVE statements

```octave
>> inc = 4.0/29.0;
>> [x,y,z] = meshgrid( -2.0:inc:2.0, -2.0:inc:2.0, -2.0:inc:2.0 );
```

\footnote{The procedure \texttt{lugen_list} invoked here is identical for both two-dimensional scalar and vector arrays and is described in Section 9.12.5.}

\footnote{A Lawrence-written OCTAVE procedure for reading files written by \texttt{lugen_list} is described in Section 9.16.5.}

\footnote{We have supposed indices starting at 1. In languages where indices start at 0, the associations would, of course, be $E(*,**,**,0)$ storing $E_x$, $E(*,**,**,1)$ storing $E_y$, and $E(*,**,**,2)$ storing $E_z$.}
>> xp = x + 1.0; xm = x - 1.0; yp = y + 1.0; ym = y - 1.0;
>> xps = xp.*xp; xms = xm.*xm; yps = yp.*yp; yms = ym.*ym;
>> zs = z.*z;
>> rmm = (xms+yms+zs).^1.5; rmp = (xms+yps+zs).^1.5;
>> rpm = (xps+yms+zs).^1.5; rpp = (xps+yps+zs).^1.5;
>> E(:,:,1) = xm.*(1.0./rmm - 1.0./rmp) + xp.*(1.0./rpp - 1.0./rpm);
>> E(:,:,2) = ym.*(1.0./rmm - 1.0./rpm) + yp.*(1.0./rpp - 1.0./rmp);
>> E(:,:,3) = z.*(1.0./rmm - 1.0./rmp + 1.0./rpp - 1.0./rpm);

The first two lines produce a grid of values for the evaluation of the vector field. Here, $x$, $y$, and $z$ each range from $-2.0$ to $2.0$, and the interval is divided into 29 segments, so all arrays except $E$ have dimensions $30 \times 30 \times 30$. (The dimensions of $E$ are $30 \times 30 \times 30 \times 3$.) The next several lines calculate a number of intermediate quantities to facilitate the final evaluation of the fields. The three lines beginning $E(:,:,\ldots)$ evaluate the $E$-field components. Finally, with the statement

>> lugen_list( E )

we write the data to a file in the desired format.\footnote{The procedure $lugen_list$ invoked here is identical for both two- and three-dimensional scalar and vector arrays and is described in Section 9.12.5.} After the filename—we here choose quadpole_octave.dat—and other labeling information has been typed in response to the sequence of prompts, the data file will be saved in the default directory.\footnote{A Lawrence-written OCTAVE procedure for reading files written by $lugen_list.m$ is described in Section 9.16.5.}

\section*{9.16 Reading Files}

Full use of the files that we have learned how to create in the previous sections, of course, depends on an ability to read files created with a program in one language into a program in another—or possibly the same—language. This section will lay out the necessary features.

\subsection*{9.16.5 \ldots with OCTAVE}

The Lawrence-generated OCTAVE function $luread_list$ parallels the OCTAVE function $lugen_list$ first described in Section 9.12.5 and provides the facility for reading into OCTAVE any file created in the format described in Section 9.3.2, whether the files were created with OCTAVE or some other program. We simply execute the statement

$Var = luread_list( \text{"FileName"} )$

where $Filename$ is the name of the file containing the data to be read (including the file type) and $Var$ is the name of the variable into which the data are to be placed when read. For example, the statement

$\gg I = luread_list( \text{"irrad_octave.dat"} )$

will recreate the variable $I$ as it was when the file $irrad_matlab.dat$ was created in Section 9.12.5. As it executes, $luread_list$ displays the descriptive lines that were placed at the beginning of the file when it was created. The procedure $luread_list.m$ that defines this function is presented in Section 9.5.
Table 9.4: Algorithm for Exercise 9.1.

PROGRAM PARK
Obtain date and time of entry to lot
Obtain date and time of exit from lot
Determine time in lot in days, hours, and minutes
Report time in lot
IF time in lot less than one day
    THEN calculate fee for less than one day
    ELSE calculate fee for one day or longer
END_IF
Report fee
END_PROGRAM

9.17 References

For OCTAVE, see Section 4.19.

9.18 Exercises

9.1. Pricing at an airport parking lot is as follows: 50 cents for the first half hour, 35 cents for the second half hour, and 25 cents for each subsequent hour (or fraction thereof) to a maximum of 250 cents per 24-hour period. The higher charge for the first hour applies only on the first day. Make the algorithm listed in Table 9.4 for determining the parking fee for a particular patron more explicit by expanding the two statements in the IF-THEN-ELSE structure. Hint: Work through several numeric examples by hand, noting particularly all decisions that you must make in order to know what arithmetic to do. Optional: Describe a procedure for determining the time in the lot from the dates and times of entry and exit. Assume first that the two dates are in the same month, but then give some thought to generalizing your procedure to handle cases where the two dates span two or more months or years.

9.2. Figure 9.5 shows three different alternative structures. Express each structure using (a) only CASE structures and (b) only IF-THEN-ELSE structures. In these figures, T, F, C, and B stand for true, false, condition, and block of statements, respectively. Use proper indentation as illustrated in the examples.

9.3. Identify the basic actions performed by the automatic pin setting apparatus at the end of a bowling alley. Then write an algorithm to control the operation of this device.

9.4. For the game of bowling, identify appropriate elementary action statements and then write an algorithm that will accept the number of pins knocked over with each ball and report the frame-by-frame scores.

9.5. (a) Cast Algorithm (6) of Section 9.2 in a form more specific to finding the largest integer in a list of integers.

(b) The algorithm shown in Table 9.5 is an alternative to the algorithm deduced in (a). Essentially, an alternative method of initializing LARGEST% is adopted. Write a few sentences identifying the advantages and disadvantages of the two different methods.

9.6. Basing your work on Algorithm (6) of Section 9.2, write an algorithm that will obtain words one at a time and ultimately report the word that would occur first if the list were to be alphabetized.
Figure 9.5: Figure for Exercise 9.2.

Table 9.5: Algorithm for Exercise 9.5.

SENTINEL% ← (agreed-upon special value)
LARGEST% ← (assumed fictitious integer known to occur before any possible real integer in the list)

LOOP
  READ ITEM%
  EXIT_LOOP WHEN ITEM% = SENTINEL%
  IF ITEM% > LARGEST%
    THEN LARGEST% ← ITEM%
  END_IF
END_LOOP
Report LARGEST%
Table 9.6: Procedure for Exercise 9.9.

PROCEDURE ???????
SCANEND% ← N%
LOOP
CARD% ← 1%
Obtain word on card CARD% and store in WORD$
LATEST_WORD$ ← WORD$
LATEST_CARD% ← CARD%
LOOP
CARD% ← CARD% + 1%
Obtain word on card CARD% and store in WORD$
IF WORD$ occurs after LATEST_WORD$
THEN BEGIN_BLOCK
LATEST_WORD$ ← WORD$
LATEST_CARD% ← CARD%
END_BLOCK
END_IF
EXIT_LOOP WHEN CARD% = SCANEND%
END_LOOP
Exchange card LATEST_CARD% with card SCANEND%
SCANEND% ← SCANEND% - 1%
EXIT_LOOP WHEN SCANEND% = 1%
END_LOOP
END_PROCEDURE

9.7. Basing your work on Algorithm (6) of Section 9.2, write an algorithm that will obtain words one at a time and ultimately report (1) the word that would appear last if the list were alphabetized, (2) the word that would appear first if the list were alphabetized, (3) the total number of words given, and (4) the position of each extreme word in the original list. Only one pass through the list is permitted.

9.8. Write an algorithm that will find and report all triplets of positive integers (zero excluded) $A\%, B\%, C\%$ satisfying $A\%*A\% + B\%*B\% = C\%*C\%$, subject to the restriction that $A\%, B\%$, and $C\%$ shall all be smaller than some value MAXNUM% supplied as input. *Hint:* Systematically examine all possibilities, but do so thoughtfully. For example, there is no point in examining cases for which $C\%<=A\%$ or $C\%<=B\%$. Express your loops so that these cases (and any others that you can reject *a priori*) will not even be considered. *Optional:* For MAXNUM% = 20%, determine the number of executions of your innermost loop.

9.9. Suppose you have $N\%$ cards laid out in a row on a table. On each card is a single word. Determine the end result of applying the mystery procedure laid out in Table 9.6 to that array of cards and choose a suitable name for the procedure.

9.10. Let the digits in an integer be counted from the left end of the integer, i.e., in the four-digit number “4358”, call “4” digit 1, “3” digit 2, “5” digit 3, and “8” digit 4. Determine the function of the mystery procedure in Table 9.7 and choose a suitable name for the procedure.

9.11. Starting with the approximation

$$ \frac{du}{dx} \bigg|_{x+\frac{1}{4}\Delta x} \approx \frac{u(x + \Delta x) - u(x)}{\Delta x} $$
Table 9.7: Procedure for Exercise 9.10.

PROCEDURE ???????
Obtain a positive integer A% from a friend
Obtain a (second) positive integer B% from a friend
CASE
  OF A% has more digits than B% DO
    LOOP
      Add digit 0 in front of B% and call result B%
      EXIT_LOOP WHEN B% and A% have same number of
digits
    END_LOOP
  OF B% has more digits than A% DO
    LOOP
      Add digit 0 in front of A% and call result A%
      EXIT_LOOP WHEN A% and B% have same number of
digits
    END_LOOP
END_CASE
WRITE A% on a piece of paper
WRITE B% under A% with corresponding digits in same
column
Draw a line under B%
DIGIT% ← number of digits in either number
CARRY% ← 0%
LOOP
  SUM% ← digit DIGIT% of A% + digit DIGIT% of B% +
  CARRY%
  IF SUM% < 10%
    THEN CARRY% ← 0%
  ELSE BEGIN_BLOCK
    CARRY% ← 1%
    SUM% ← SUM% - 10%
  END_BLOCK
END_IF
WRITE SUM% under digit DIGIT% of B%
DIGIT% ← DIGIT% - 1%
EXIT_LOOP WHEN DIGIT% = 0%
END_LOOP
IF CARRY% = 1%
  THEN WRITE CARRY% in front of all digits beneath line
END_IF
END_PROCEDURE

deduce the finite difference approximation

\[ \frac{d^2u}{dx^2} \bigg|_x \approx \frac{u(x + \Delta x) - 2u(x) + u(x - \Delta x)}{\Delta x^2} \]

for the second derivative.

9.12. Write, compile, and test a program that asks for the input of a temperature in Celsius and prints out the corresponding temperature in Fahrenheit. To make it a bit more of a challenge, write the
Consider two circular current loops, each of radius \( a \) and lying with its center on and its plane perpendicular to the \( z \) axis. The first loop is centered at the point \((0, 0, b)\) and the second loop is centered at the point \((0, 0, -b)\). The axial component of the magnetic field at the point \((0, 0, z)\) is given by the equation

\[
B(z) = \frac{1}{2} B_0 \left( a^2 + b^2 \right)^{3/2} \left[ \frac{1}{a^2 + (z + b)^2} + \frac{1}{a^2 + (z - b)^2} \right]^{3/2}
\]

where \( B_0 \) is the magnetic field at the origin. This field can be considered as a function not only of \( z \), the coordinate of a point on the \( z \) axis, but also of \( b \), (half) the separation of the two loops. Create a file conforming to the structure described in Section 9.3.2 and containing values of this function seen as a two-dimensional scalar function of \( z/a \) and \( b/a \). 

\textit{Suggestion:} Write values of \( B/B_0 \) into the file.

The trajectory of a particle in three-dimensional space is given parametrically as a function of time \( t \) by the position vector

\[
r = x(t) \, \hat{i} + y(t) \, \hat{j} + z(t) \, \hat{k}
\]

You desire to fathom out the general character of this trajectory by using a graphical visualization tool that does not have much computational capability. Thus, you must generate the data using one tool but will visualize the trajectory with another tool. You elect to use an ASCII file to communicate the data from the first tool to the second. Suppose that the ASCII file produced by the first tool is to be structured as follows:

- five lines of text describing the contents of the file and its origin,
- one line containing the number of points \( N \) on the trajectory included in the file, and
- \( N \) lines, each of which contains four floating point values separated by commas, those values being in order \( t, x(t), y(t), \) and \( z(t) \) for a point on the trajectory. (The \( N \) lines are ordered by increasing value of \( t \)).

Describe a general procedure to create this file and then implement that procedure in at least one language of your choice, testing your program(s) with the trajectory given by

\[
r = \cos t \, \hat{i} + \sin t \, \hat{j} + 0.1t \, \hat{k}
\]

which describes the path followed by a charged particle in a constant magnetic field along the \( z \) axis.

Following the pattern illustrated in Section 9.13, use at least one language to create a file containing values for at least one of the three-dimensional scalar fields

\[
p_{3,1,0}(x, y, z) = \frac{8}{(27)^2 \pi} \rho^2 \left( 1 - \frac{\rho}{6} \right)^2 e^{-2\rho/3} \cos^2 \theta
\]

\[
p_{3,1,1}(x, y, z) = \frac{4}{(27)^2 \pi} \rho^2 \left( 1 - \frac{\rho}{6} \right)^2 e^{-2\rho/3} (1 - \cos^2 \theta)
\]

\[
p_{3,2,1}(x, y, z) = \frac{3}{(27)^3 \pi} \rho^4 e^{-2\rho/3} \cos^2 \theta (1 - \cos^2 \theta)
\]

\[
p_{3,2,2}(x, y, z) = \frac{3}{4(27)^3 \pi} \rho^4 e^{-2\rho/3} (1 - \cos^2 \theta)^2
\]

giving the probability density for the hydrogen states \((n, l, m) = (3, 1, 0), (n, l, m) = (3, 1, 1), (n, l, m) = (3, 2, 1), \) and \((n, l, m) = (3, 2, 2)\). These fields are expressed in dimensionless form, where \( \rho \) is the radial coordinate in units of the Bohr radius. In terms of the Cartesian coordinates \( x, y, z \), \( \rho = \sqrt{x^2 + y^2 + z^2} \) and \( \cos \theta = z/\rho \). 

\textit{Hint:} To avoid divisions by zero, recast the expressions in terms of \((x, y, z)\) explicitly before evaluating any of them.
9.18. The (gauge) pressure \( p(x, y, z, t) \) inside a cubical box located in the region \( 0 \leq x, y, z \leq a \) is given by

\[
p(x, y, z, t) = A \sin \left( \frac{l \pi x}{a} \right) \sin \left( \frac{m \pi y}{a} \right) \cos \left( \frac{n \pi z}{a} \right) \cos \omega l m n t
\]

where \( l, m, \) and \( n \) are positive integers. Using at least one language, create files containing the pressure distribution inside the box at \( t = 0 \) for several different values of \( l, m, \) and \( n \).

9.19. A point charge of strength \( q \) is located on the \( y \) axis at \( r_+ = a \hat{j} \) and a point charge of strength \(-q\) is located on the \( y \) axis at \( r_- = -a \hat{j} \). The electric field \( \mathbf{E} \) at the point \( r = x \hat{i} + y \hat{j} \) in the \( xy \) plane is given in mks units by

\[
\mathbf{E}(x, y) = \frac{q}{4 \pi \epsilon_0} \left[ \frac{x \hat{i} + (y - a) \hat{j}}{[x^2 + (y - a)^2]^{3/2}} - \frac{x \hat{i} + (y + a) \hat{j}}{[x^2 + (y + a)^2]^{3/2}} \right]
\]

Expressing coordinates in terms of the dimensionless variables \( \tilde{x} = x/a \) and \( \tilde{y} = y/a \) and measuring \( \mathbf{E} \) in the unit \( q/(4 \pi \epsilon_0 a^2) \), create a file conforming to the structure described in Section 9.3.2 and containing the \( x \) and \( y \) components of this field over the interval \(-2a \leq x, y \leq 2a\). Divide the interval in each coordinate direction into about 25 segments, but choose the precise number carefully so as to avoid division by zero in evaluating the field at any point. Optional: Read your file into a suitable graphical display program and produce graphs showing the character of this field.

9.20. The velocity field of an object rotating about the \( z \) axis with angular momentum \( \omega \) is given in terms of the angular velocity and the position vector \( \mathbf{r} \) by the expression

\[
\mathbf{v} = \omega \times \mathbf{r} = \omega \mathbf{k} \times (x \hat{i} + y \hat{j} + z \hat{k}) = \omega (-y \hat{i} + x \hat{j})
\]

Choosing a unit of length \( a \) and expressing coordinates in units of \( a \) and velocities in units of \( wa \), create a file conforming to the structure described in Section 9.3.2 and containing the \( x \) and \( y \) components of this field over the interval \(-2a \leq x, y \leq 2a\). Choose the number of divisions in each coordinate direction so as to generate a display which is neither too sparse to be useful nor too nor too busy to be intelligible. Optional: Read your file into a suitable graphical display program and produce graphs showing the character of this field.

9.21. A point charge of strength \( q \) is located on the \( z \) axis at \( r_+ = a \hat{k} \) and a point charge of strength \(-q\) is located on the \( z \) axis at \( r_- = -a \hat{k} \). The electric field \( \mathbf{E} \) at the point \( r = x \hat{i} + y \hat{j} + z \hat{k} \) is given in mks units by

\[
\mathbf{E}(x, y, z) = \frac{q}{4 \pi \epsilon_0} \left[ \frac{x \hat{i} + y \hat{j} + (z - a) \hat{k}}{[x^2 + y^2 + (z - a)^2]^{3/2}} - \frac{x \hat{i} + y \hat{j} + (z + a) \hat{k}}{[x^2 + y^2 + (z + a)^2]^{3/2}} \right]
\]

Expressing coordinates in terms of the dimensionless variables \( \tilde{x} = x/a \), \( \tilde{y} = y/a \), and \( \tilde{z} = z/a \) and measuring \( \mathbf{E} \) in the unit \( q/(4 \pi \epsilon_0 a^2) \), create a file conforming to the structure described in Section 9.3.2 and containing the \( x \), \( y \), and \( z \) components of this field over the interval \(-2a \leq x, y, z \leq 2a\). Divide the interval in each coordinate direction into about 25 segments, but choose the precise number carefully so as to avoid division by zero in evaluating the field at any point. Optional: Read your file into a suitable graphical display program and produce graphs showing the character of this field.

9.22. Write and test a program to ask for the latitude and longitude of both a point of departure \( D \) and a point of arrival \( A \) on the surface of the earth and then calculate and print out the "crow flies" distance along a great circle route from \( D \) to \( A \). Make sure your program prints the shorter of the two distances, regardless of the location of the points, and make sure your program doesn’t run into difficulties if the two points happen to be at opposite ends of a diameter. Take the earth to be a perfect sphere with a circumference of 24900 miles (radius 3963 miles). For purposes of testing, note that Albany, NY, is at \([43^\circ 40'N, 73^\circ 45'W]\); Grand Junction, CO, is at \([39^\circ 55'N, 108^\circ 33'W]\); Los Angeles, CA, is at \([34^\circ 3'N, 118^\circ 15'W]\); Appleton, WI, is at \([44^\circ 16'N, 88^\circ 25'W]\); Calcutta India, is at \([22^\circ 32'N, 88^\circ 20'W]\); Sydney, Australia, is at \([33^\circ 52'S, 151^\circ 12'E]\); Paris, France, is at \([48^\circ 49'N, 2^\circ 29'E]\); and Stockholm, Sweden, is at \([59^\circ 21'N, 18^\circ 4'E]\).
function [] = lugen_list(array)
% LUGEN_LIST writes data in an array to a file
% LUGEN_LIST facilitates writing data stored in an array in
% OCTAVE into an ASCII file in a standard format for purposes
% both of saving the data and of transferring it to other
% programs for further examination. The array to be written
% to the file is communicated to LUGEN_LIST through its one
% argument.

% ***** Fetch file name and labeling information *****
fprintf(1, '\nAll entries are strings and must be enclosed\n')
fprintf(1, 'in single quotation marks.\n')
FileName = input( 'File name [untitled.dat]: ' );
if isempty(FileName) FileName = 'untitled.dat'; end;
title = input( 'Title : ' );
author = input( 'Author : ' );
date = input( 'Date : ' );
comm1 = input( 'Comment 1 [**]: ' );
if isempty(comm1) comm1 = '**'; end;
comm2 = input( 'Comment 2 [**]: ' );
if isempty(comm2) comm2 = '**'; end;
comm3 = input( 'Comment 3 [**]: ' );
if isempty(comm3) comm3 = '**'; end;
comm4 = input( 'Comment 4 [**]: ' );
if isempty(comm4) comm4 = '**'; end;

% ***** Concatenate title, author and date *****
% ***** Find dimensions of array *****
title = [title,'; Author: ',author,'; Date: ',date];
[xdim,ydim,zdim,narr] = size(array);

% ***** Open file, write labeling lines, write dimensions *****
id = fopen( FileName, 'w' );
fprintf( id, '%s
', title );
fprintf( id, '%s
', comm1 );
fprintf( id, '%s
', comm2 );
fprintf( id, '%s
', comm3 );
fprintf( id, '%s
', comm4 );
fprintf( id, '%d
', narr );
fprintf( id, '%d
', xdim );
fprintf( id, '%d
', ydim );
fprintf( id, '%d
', zdim );

% ***** Write data in array to file *****
for n = 1:narr
    for k = 1:zdim
        for j = 1:ydim
            for i = 1:xdim


fprintf(id, '%g\n', array(i,j,k,n));
end
end
end

% ***** Close file; print message to screen *****

status = fclose(id);
fprintf(1, ['\nFile ''',FileName,''', saved successfully!\n'] )
function U = luread_list( FileName )
% LUREAD_LIST reads files of the proper structure into MATLAB.
% LUREAD_LIST reads a file created in the format described in
% CPSUP and stores the array internal to MATLAB. It is invoked
% with the command
%  
%   U = luread_list( 'FileName' )
%  
% where 'filename' specifies the file and U identifies the name of the
% variable into which the data will be read.

% ***** Print starting message, open file, read/display headers *****
fprintf( 1, '\nReading File\n' )
id = fopen( FileName, 'r' );
head1 = fgetl( id ); fprintf( 1, [head1,'
' ] )
head2 = fgetl( id ); fprintf( 1, [head2,'
' ] )
head3 = fgetl( id ); fprintf( 1, [head3,'
' ] )
head4 = fgetl( id ); fprintf( 1, [head4,'
' ] )
head5 = fgetl( id ); fprintf( 1, [head5,'
' ] )

% ***** Read dimensions *****
narr = fscanf( id, '\%d', 1 );
xdim = fscanf( id, '\%d', 1 );
ydim = fscanf( id, '\%d', 1 );
zdim = fscanf( id, '\%d', 1 );

% ***** Create U, read file into U, and close file *****
U = zeros( xdim, ydim, zdim, narr );
for n = 1:narr
    for k = 1:zdim
        for j = 1:ydim
            for i = 1:xdim
                U(i,j,k,n) = fscanf( id, '\%f', 1 );
            end
        end
    end
end
status = fclose( id );

%***** Print reassuring message *****
fprintf( 1, [ '\nFile ', FileName, ' read successfully!\n' ] )
Chapter 11

Solving Ordinary Differential Equations

Many fundamental laws of physics relate the rate at which the physical properties of a system change to the properties themselves. These physical laws lead inevitably to differential equations satisfied by the quantities describing the system. While some of these equations admit closed form, symbolic solutions, most can be solved only through numerical approximation. We begin this chapter by identifying several physical situations, the full addressing of which requires us to solve an ordinary differential equation (ODE) or a coupled set of such equations. Then we illustrate how to use symbolic algebra systems to approach those that can be solved analytically, describe a few of many available numerical algorithms (with attention to their accuracy), and—finally—describe ways to solve representative ODEs using a variety of numerical approaches and computational tools.

Differential equations fall into many, sometimes overlapping, categories. We limit ourselves in this chapter to ordinary differential equations, which involve only one independent variable. Most equations of interest in physics are first-order (containing no derivatives higher than the first) or second-order (containing no derivatives higher than the second), but occasionally higher order equations may arise. Whatever their order, these equations may be linear (each term depending on the dependent variable only through either its first power or the first power of one of its derivatives) or non-linear (at least one term violating the constraint in the previous parentheses). They may be homogeneous (no term free of the dependent variable) or inhomogeneous (at least one term free of the dependent variable). The coefficients may be constant or may depend on the independent variable. Most will contain parameters characterizing the system of interest, though recasting the original equations in dimensionless form may reduce the number of distinct parameters—or even eliminate them altogether. We may be confronted with a single equation (one dependent variable) or with a system of equations (two or more dependent variables), and the members of the system may be coupled (more than one of the dependent variables appearing in at least one of the equations) or decoupled (no member of the system containing more than one of the dependent variables).

For complete statement of a problem, the applicable ODEs must be supplemented with auxiliary conditions, the number of which equals the sum of the orders of the equations at hand. A single first-order equation requires one condition; stipulation of the value of the dependent variable at a specific value of the independent variable is sufficient to select a unique solution from the family of solutions defined by the differential equation alone. A single second-order equation requires two conditions but, in this case, we have some choices. We might, for example, stipulate the value of the dependent variable and the value of its first derivative at a single value of the independent variable, e.g., position and velocity at an initial time. In that case, we would be dealing with an initial value problem (IVP). Alternatively, we might stipulate the value of the dependent variable at each of two different values of the independent variable, e.g., displacement of a string at each of its two ends. In that case, we would be dealing with a boundary value problem (BVP).
The approach to solving a particular ODE or system of ODEs may well be dictated by the category into which the equation or equations fall. The approach will also be influenced by whether we are dealing with an IVP or a BVP. The examples chosen for this chapter illustrate several of these situations.

11.1 Sample Problems

In this section, we identify several physical contexts that lead to differential equations, and we determine the specific differential equation arising in each case. Solutions to the resulting equations by a variety of symbolic and numerical means will be explored in the remainder of this chapter.

In most cases the statement of the problem of interest will contain several constants or parameters. Some reside in the ODE itself while others reside in the initial or boundary conditions. The presence of such constants gives rise to two complications. First, a system of equations containing many constants is much more difficult to explore than a system containing only a few constants. Second, in some cases, the values of the constants will be either very large (e.g., planetary distances or masses) or very small (e.g., atomic distances or masses). In these cases, finding appropriate initial conditions can be difficult. Additionally, numbers of these magnitudes can potentially cause floating point overflow or underflow. Frequently, both of these complications can be made less severe by casting the differential equation(s) and associated initial or boundary conditions in dimensionless form. To accomplish that objective, we begin by choosing judicious, non-standard units in terms of which to express the independent and dependent variables. Then, we rescale these variables to express them in the chosen units. Sometimes, all parameters in the equations and the initial or boundary conditions will disappear; more often, a small number of (dimensionless) combinations of parameters will remain. In any case, the recast problem is almost certain to be simpler than the original problem, partly because the solution depends on fewer “real” parameters and partly because the significant values of all quantities are likely to have order of magnitude one. Anticipating that dimensionless presentations will facilitate some of our subsequent solutions, we shall conclude several of the subsections in this section by illustrating how the strategy described in this paragraph would be implemented for the equations in those subsections.

11.1.1 Projectile in a Viscous Medium

The projectile shown in Fig. 11.1 moves in a viscous medium near the surface of the earth. It experiences two forces, the gravitational attraction of the earth\(^1\) \(-mg \hat{k}\) and the viscous force \(F_v\), from the medium in which it moves. The former is directed downward and the latter is directed

\(^1\)We choose a coordinate system in which positive \(z\) is directed upward, and we take \(m\) and \(g\) to be positive.
11.1. SAMPLE PROBLEMS

opposite to the velocity $\mathbf{v}$. Usually, the magnitude of the viscous force is a function of the speed with which the projectile moves, symbolically $|\mathbf{F}_v| = f(|\mathbf{v}|)$. In general, if the projectile has mass $m$, Newton’s second law yields the equation of motion

$$m \frac{d^2 \mathbf{r}}{dt^2} = -mg \hat{k} + \mathbf{F}_v(|\mathbf{v}|) = -mg \hat{k} - f(|\mathbf{v}|) \frac{\mathbf{v}}{|\mathbf{v}|} \quad (11.1)$$

which we are to solve subject to the general initial conditions,

$$\mathbf{r}(0) = \mathbf{r}_0 = x_0 \hat{i} + y_0 \hat{j} + z_0 \hat{k} \quad ; \quad \mathbf{v}(0) = \mathbf{v}_0 = v_{x_0} \hat{i} + v_{y_0} \hat{j} + v_{z_0} \hat{k} \quad (11.2)$$

Of course, we also need to know the precise dependence of the function $f(|\mathbf{v}|)$ on the speed of the projectile. In the simplest case (when the speed of the projectile is small enough), $f$ is simply linearly proportional to that speed, $f(|\mathbf{v}|) = b |\mathbf{v}|$ ($b$ a positive constant), and the equation of motion reduces to

$$m \frac{d^2 \mathbf{r}}{dt^2} = -mg \hat{k} - b\mathbf{v} = -mg \hat{k} - b \frac{d\mathbf{r}}{dt} \quad (11.3)$$

or, in component form, to

$$m \frac{d^2 x}{dt^2} = -b \frac{dx}{dt} \quad ; \quad m \frac{d^2 y}{dt^2} = -b \frac{dy}{dt} \quad ; \quad m \frac{d^2 z}{dt^2} = -mg - b \frac{dz}{dt} \quad (11.4)$$

This system of equations is uncoupled, since each of the three independent variables $x$, $y$, and $z$ satisfies its own private equation that does not involve either of the other variables. They are second-order and linear, and the coefficients are constant. The first two are homogeneous and, because of the term $-mg$, the third is inhomogeneous. The equations involve the parameters $m$, $b$, and $g$, and their solutions will depend on these parameters and on the six initial values in Eq. (11.2).

When the speed is too large for a linear approximation to the viscous damping, we can sometimes take the viscous force to be proportional instead to the square of the speed, $f = b|\mathbf{v}|^2$ ($b$ a positive constant, though not the same constant as in the previous paragraph). This time, the equation of motion reduces to

$$m \frac{d^2 \mathbf{r}}{dt^2} = -mg \hat{k} - b|\mathbf{v}|\mathbf{v} = -mg \hat{k} - b \frac{d\mathbf{r}}{dt} \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2} \quad (11.5)$$

or, in component form, to

$$m \begin{bmatrix} \frac{d^2 x}{dt^2}, \frac{d^2 y}{dt^2}, \frac{d^2 z}{dt^2} \end{bmatrix} = [0, 0, -mg] - b \begin{bmatrix} \frac{dx}{dt}, \frac{dy}{dt}, \frac{dz}{dt} \end{bmatrix} \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2} \quad (11.6)$$

This system of equations is second-order and distinctly non-linear, and its members are coupled because each of the equations involves all three of the dependent variables. Even if the motion occurs in only the vertical dimension (projectile tossed straight up or, simply, released from rest and allowed to drop), the equation, which then is

$$m \frac{d^2 z}{dt^2} = -mg - b \frac{dz}{dt} \left| \frac{dz}{dt} \right| \quad (11.7)$$

is complicated by the absolute value in the viscous term. These equations involve the parameters $m$, $b$, and $g$, and their solutions will depend on these parameters and on the six initial values in Eq. (11.2).

To cast these equations in a dimensionless form, we would start by choosing a unit of length, say $\ell$—which might be chosen arbitrarily, chosen to be one of the initial positions, or merely symbolized initially and chosen later to simplify the equations as we converge on their dimensionless form. Then, dividing by $m$ and rescaling the dimensional coordinates by introducing the dimensionless
coordinates \( \mathbf{x} = x/a, \mathbf{y} = y/a, \) and \( \mathbf{z} = z/a \) (i.e., \( \mathbf{r} = r/a \)), we find that Eqs. (11.3) and (11.5) become
\[
\frac{d^2 \mathbf{r}}{dt^2} = -\frac{g}{a} \mathbf{k} - \frac{b}{m} \frac{d\mathbf{r}}{dt} \tag{11.8}
\]
and
\[
\frac{d^2 \mathbf{r}}{dt^2} = -\frac{g}{a} \mathbf{k} - \frac{ba}{m} \frac{d\mathbf{r}}{dt} \left( \left( \frac{d\mathbf{x}}{dt} \right)^2 + \left( \frac{d\mathbf{y}}{dt} \right)^2 + \left( \frac{d\mathbf{z}}{dt} \right)^2 \right) \tag{11.9}
\]
respectively. Next, we recognize that \( \sqrt{g/a} \) is dimensionally a frequency.\(^2\) Hence, the variable \( \tau = t \sqrt{g/a} \) provides a suitable rescaling of the independent variable. In terms of \( \tau \), Eqs. (11.8) and (11.9) become
\[
\frac{d^2 \mathbf{r}}{d\tau^2} = -\mathbf{k} - \frac{b}{m} \sqrt{\frac{a}{g}} \frac{d\mathbf{r}}{d\tau} \tag{11.10}
\]
and
\[
\frac{d^2 \mathbf{r}}{d\tau^2} = -\mathbf{k} - \frac{ba}{m} \frac{d\mathbf{r}}{d\tau} \sqrt{\left( \frac{d\mathbf{x}}{d\tau} \right)^2 + \left( \frac{d\mathbf{y}}{d\tau} \right)^2 + \left( \frac{d\mathbf{z}}{d\tau} \right)^2} \tag{11.11}
\]
Finally, we introduce the symbol \( \beta \) for the dimensionless quantity \( b \sqrt{a/g}/m \) in the first of these equations and for the dimensionless quantity \( ba/m \) in the second.\(^3\) In the end, the equations in dimensionless form are
\[
\frac{d^2 \mathbf{r}}{d\tau^2} = -\mathbf{k} - \beta \frac{d\mathbf{r}}{d\tau} \tag{11.12}
\]
and
\[
\frac{d^2 \mathbf{r}}{d\tau^2} = -\mathbf{k} - \beta \frac{d\mathbf{r}}{d\tau} \sqrt{\left( \frac{d\mathbf{x}}{d\tau} \right)^2 + \left( \frac{d\mathbf{y}}{d\tau} \right)^2 + \left( \frac{d\mathbf{z}}{d\tau} \right)^2} \tag{11.13}
\]
These equations are, of course, to be solved subject to the dimensionless initial conditions
\[
\mathbf{r}(0) = \frac{\mathbf{r}_0}{a}; \quad \mathbf{v}(0) = \frac{d\mathbf{r}}{dt}(0) = \frac{d(r/a)}{d(t \sqrt{g/a})}(0) = \frac{\mathbf{v}_0}{\sqrt{g/a}} \tag{11.14}
\]
from which we conclude that dimensionless velocities are measured in units of \( \sqrt{ga} \).

In truth, however, these particular equations are not really second-order equations. They can be reduced to first-order equations by focusing attention on the components \( [v_x, v_y, v_z] \) of the velocity as the dependent variables. In those terms, the three members of Eq. (11.4) become
\[
m \frac{dv_x}{dt} = -bv_x; \quad m \frac{dv_y}{dt} = -bv_y; \quad m \frac{dv_z}{dt} = -mg - bv_z \tag{11.15}
\]
the three members of Eq. (11.6) become
\[
m \begin{bmatrix} \frac{dv_x}{dt}, \frac{dv_y}{dt}, \frac{dv_z}{dt} \end{bmatrix} = \begin{bmatrix} 0, 0, -mg \end{bmatrix} - b [v_x, v_y, v_z] \sqrt{v_x^2 + v_y^2 + v_z^2} \tag{11.16}
\]
and Eq. (11.7) becomes
\[
m \frac{dv_z}{dt} = -mg - bv_z |v_z| \tag{11.17}
\]
Once these first-order equations have been solved, the components of the position vector can then be found by solving the differential equations
\[
\frac{dx}{dt} = v_x; \quad \frac{dy}{dt} = v_y; \quad \frac{dz}{dt} = v_z \tag{11.18}
\]
\(^2\)Using \([\ldots]\) to indicate “the dimensions of \( \ldots \)”, we argue \( [g] = \text{length/time}^2, [a] = \text{length} \Rightarrow [g/a] = \text{time}^{-2} \Rightarrow [\sqrt{g/a}] = \text{time}^{-1} \).
\(^3\)Remember that \( b \) is not the same in the two instances.
\(^4\)Here (and in all subsequent cases where a chosen unit is identified but not checked), you should take a moment to verify that the identified unit has the proper dimensions.
11.1 SAMPLE PROBLEMS

11.1.2 Chain Radioactive Decay

A wholly different context in which systems of ODEs arise lies in radioactive decay. The fundamental law asserts that a sample of a particular radioisotope decays at a rate proportional to the quantity of (undecayed) material present in the sample. Thus, for the decay chain shown in Fig. 11.2, we would write the system of three differential equations

\[
\begin{align*}
\frac{dA}{dt} &= -k_A A ; \\
\frac{dB}{dt} &= k_A A - k_B B ; \\
\frac{dC}{dt} &= k_B B
\end{align*}
\]  

(11.19)

where \( k_A \) and \( k_B \) are decay constants (parameters); \( A(t) \), \( B(t) \), and \( C(t) \) are the number of nuclei of each species present; and nucleus C is assumed to be stable. These equations are linear, first-order, and homogeneous, and they have constant coefficients. They are, however, coupled, since each of the second and third of them involves two of the dependent variables. They also support a conservation law: adding the three equations yields that

\[
\frac{d}{dt}(A + B + C) = 0 = \Rightarrow A + B + C = \text{constant} \tag{11.20}
\]

As always we, of course, need initial values, e.g., \( A(0) = A_0 \), \( B(0) = 0 \), and \( C(0) = 0 \), before the differential equations have a unique solution, and that solution will depend on the parameters \( k_A \) and \( k_B \) and on the three initial values.

To cast these equations in a dimensionless form, we choose a reference amount—here, conveniently, \( A_0 \), the initial amount of \( A \)—as the unit for measuring the quantities of \( A \), \( B \), and \( C \). Then, we rescale the values for these quantities by introducing the dimensionless variables \( \overline{A} = A/A_0 \), \( \overline{B} = B/A_0 \), and \( \overline{C} = C/A_0 \). Next, dividing the equations by \( k_A A_0 \) and introducing the dimensionless quantities \( \overline{t} = k_A t \) and \( \overline{k} = k_B/k_A \), we conclude that

\[
\begin{align*}
\frac{d\overline{A}}{d\overline{t}} &= -\overline{A} ; \\
\frac{d\overline{B}}{d\overline{t}} &= \overline{A} - \overline{k} \overline{B} ; \\
\frac{d\overline{C}}{d\overline{t}} &= \overline{k} \overline{B}
\end{align*}
\]  

(11.21)

with the initial conditions \( \overline{A}(0) = 1 \), \( \overline{B}(0) = 0 \), and \( \overline{C}(0) = 0 \), where we now regard the dependent variables as functions of the dimensionless time \( \overline{t} \). In short, we discover that this problem possesses only one “real” parameter \( \overline{k} \). Only the ratio of the rate constants conveys any significant distinction among different realizations of this decay. Everything else is simply a matter of scaling, either on the time variable or on the dependent variables as a group. The essential physics is both easier to explore and easier to comprehend when the problem is viewed from a dimensionless perspective.

---

5We shall later see that conservation laws can sometimes prove valuable in assessing the accuracy of solutions.
11.1.3 Exponential and Logistic Growth

An important illustration of a non-linear first-order equation occurs in population biology, where—in the absence of predation—the population of a species grows at a rate proportional to that population until the population becomes so large that individual organisms compete significantly with one another for space and/or food. In a simple model, the effects of competition are proportional to the likelihood that one organism will encounter another—a likelihood that is proportional to the square of the population. Thus, a population subject to both effects will evolve in accordance with the first-order, non-linear equation

\[
\frac{dN}{dt} = kN \left(1 - \frac{N}{N_c}\right) \quad (11.22)
\]

where \(N(t)\) is the population, \(k\) is the *growth rate*, and \(N_c\), which is the value of \(N\) at which its rate of growth becomes zero, is the *carrying capacity* of the environment. If, in particular, \(N(t)\) ever equals \(N_c\), then \(dN/dt = 0\) and \(N\) ceases to change; the population will attain an equilibrium, which it maintains forever after. We must, of course, know the initial population \(N(0) = N_0\) before a complete solution to this equation can be found, and that solution will depend on the parameters \(k\) and \(N_c\) and on the initial value \(N_0\).

Two different terms are used to label the solutions to Eq. (11.22). If \(N_c\) is infinite (or, more realistically, \(N(t) \ll N_c\)), then the second term in the parentheses on the right is negligible and the resulting growth is said to be exponential, though the growth will actually be a decay if \(k < 0\). If, on the other hand, \(N(t)\) is not small compared to \(N_c\), both terms are important, the exponential growth of the first case reaches a ceiling, and the growth is said to be logistic.\(^6\)

The dimensionless version of this equation is quickly found. We choose \(N_c\) as the reference population, introduce the dimensionless population \(\overline{N} = N/N_c\) and the dimensionless time \(\overline{t} = kt\), and find that the equation and initial condition reduce to

\[
\frac{d\overline{N}}{d\overline{t}} = \overline{N}(1 - \overline{N}) \quad ; \quad \overline{N}(0) = \frac{N_0}{N_c} \quad (11.23)
\]

All parameters disappear from the differential equation but the initial population—now measured in units of the carrying capacity—remains as a single parameter in the problem.

11.1.4 Forced, Driven, Damped Harmonic Oscillation

A particularly important, second-order differential equation arises in several contexts. Suppose, for example, as shown in Fig. 11.3, an object of mass \(m\) moves on a horizontal, frictionless surface under the action of forces applied by a Hooke’s law spring of constant \(k\), a viscous shock absorber of damping constant \(b\), and an externally applied time-dependent force \(F(t)\). Newton’s second law leads us to write the linear, second-order, constant coefficient equation of motion governing this system as

\[
md\frac{d^2x}{dt^2} = -kx - b\frac{dx}{dt} + F(t) \quad (11.24)
\]

Here, \(x\) is measured from the position of the object when the spring is neither stretched nor compressed. This equation is inhomogeneous if \(F \neq 0\) and homogeneous if \(F = 0\). As always, we require initial conditions, which will have the general form

\[
x(0) = x_0 \quad ; \quad v(0) = v_0 \quad (11.25)
\]

before the problem is fully stated.

\(^6\)Logistic growth is also sometimes said to follow a sigmoid curve because of the stylized ‘S’ shape that solutions exhibit when the initial population is much smaller than the carrying capacity. Graphs of this shape will be found in subsequent sections of this chapter.
To cast this equation in dimensionless form, we introduce a unit of length, say $\ell$, rescale position with the expression $\bar{x} = x/\ell$, introduce a unit of time, say $\tau$, rescale the physical time with the expressions $\bar{t} = t/\tau$, and thereby transform Eq. (11.24) to

$$\frac{d^2\bar{x}}{d\bar{t}^2} = -\frac{\tau^2 k}{m} \bar{x} - \frac{\tau b}{m} \frac{d\bar{x}}{d\bar{t}} + \frac{\tau^2 F(\tau t)}{m\ell} \tag{11.26}$$

The term on the left in this equation is now dimensionless. Thus, all terms on the right must be dimensionless as well. In particular, the combination $\tau^2 k/m$ must be dimensionless. Remember, however, that $\tau$ is at the moment merely a symbol; we have not yet made a well defined choice for the unit of time, and we are free to choose $\tau$ any way we like. Clearly, the choice $\tau^2 k/m = 1$ or $\tau = \sqrt{m/k}$ is judicious. With this choice, the equation of motion becomes

$$\frac{d^2\bar{x}}{d\bar{t}^2} = -\bar{x} - \frac{b}{\sqrt{mk}} \frac{d\bar{x}}{d\bar{t}} + \bar{F}(\bar{t}) \quad \Rightarrow \quad \frac{d^2\bar{x}}{d\bar{t}^2} = -\bar{x} - \beta \frac{d\bar{x}}{d\bar{t}} + \bar{F}(\bar{t}) \tag{11.27}$$

where, in the final form, we have set $\beta = b/\sqrt{mk}$ and $\bar{F}(\bar{t}) = F(\tau t)/k\ell$. The single, dimensionless parameter $\beta$ contains the essential influence of the three parameters $m$, $b$, and $k$ once differences attributable to scaling have been removed. The dimensionless force $\bar{F}$ expresses the physical force in units of $k\ell$, which—note—is the force that the spring would exert if extended by the chosen unit of length!

We must, of course, also translate the initial conditions of Eq. (11.25) into dimensionless form, finding that

$$\bar{x}(0) = \frac{x_0}{\ell} \quad ; \quad \bar{v}(0) = \frac{d\bar{x}}{d\bar{t}}(0) = \frac{d(x/\ell)}{d(t/\tau)} = \frac{v_0}{\ell/\tau} \tag{11.28}$$

and we conclude that dimensionless velocities will be measured in units of $\ell/\tau$, which is the speed of an object that moves the reference distance $\ell$ in the reference time $\tau$.

Alternatively (and, in some approaches to solution, necessarily), we would recast this single second-order differential equation as a pair of first-order equations, either

$$\frac{dx}{dt} = v \quad ; \quad m\frac{dv}{dt} = -kx - bv + F(t) \tag{11.29}$$

in the original dimensional form, or

$$\frac{d\bar{x}}{d\bar{t}} = \bar{v} \quad ; \quad \frac{d\bar{v}}{d\bar{t}} = -\bar{x} - \beta \bar{v} + \bar{F}(\bar{t}) \tag{11.30}$$

in dimensionless form. All dimensionless quantities are those introduced earlier in this section.
11.1.5 An LRC Resonant Circuit

An equation mathematically identical in form to Eqs. (11.29) and (11.30) arises in the description of a series RLC circuit excited by a signal generator, as shown in Fig. 11.4. If we take positive current \( i(t) \) to flow clockwise and understand that \( q(t) \) represents the charge on the left plate of the capacitor, then Kirchhoff’s loop equation and the properties of the several components lead to the equation

\[
V(t) - L \frac{di}{dt} - \frac{q}{C} - iR = 0 \implies L \frac{di}{dt} + iR + \frac{q}{C} = V(t) \tag{11.31}
\]

As it stands, this equation looks to be first order but it involves two variables \( i(t) \) and \( q(t) \). We can complete the statement of a problem having a unique solution by recognizing the relationship

\[
i = \frac{dq}{dt} \tag{11.32}
\]

between \( i \) and \( q \) and supplementing what is now a pair of coupled first-order, linear, constant coefficient, inhomogeneous equations with the general initial conditions

\[
q(0) = q_0; \quad \frac{dq}{dt}(0) = i(0) = i_0 \tag{11.33}
\]

Alternatively, we could substitute Eq. (11.32) into Eq. (11.31) to find the equivalent, single, second-order, linear, constant coefficient, inhomogeneous equation

\[
L \frac{d^2q}{dt^2} + R \frac{dq}{dt} + \frac{q}{C} = V(t) \tag{11.34}
\]

The recasting of these equations in dimensionless form is left as an exercise.

Adopting the correspondences, \( q \leftrightarrow x, L \leftrightarrow m, R \leftrightarrow b, 1/C \leftrightarrow k, V(t) \leftrightarrow F(t), \) and \( t \leftrightarrow t \), we can turn Eq. (11.34) into Eq. (11.24). Thus, mathematically, the driven, damped mechanical oscillator and the RLC circuit exhibit exactly analogous behavior, and the behavior of the RLC circuit simulates the behavior of the mechanical oscillator.\(^7\)

11.1.6 Coupled Oscillators

Consider next the system shown in Fig. 11.5 consisting of two objects of equal mass \( m \) connected along a line, each to a fixed wall by springs of constant \( k \) and each to the other by a coupling spring.

\(^7\)Years ago, when digital computers were not as fast as they have come to be and smooth graphical output from a digital computer was unheard of, correspondences such as this one were the basis of the analog computer, whereon we could easily set up electronic circuits whose behavior simulated the behavior of more expensive mechanical systems. With an analog computer, we could learn about mechanical systems by observing the real-time variation of the voltages and currents at various points in an analogous electronic circuit.
11.1. SAMPLE PROBLEMS

Figure 11.5: A system of two coupled objects.

Figure 11.6: An object moving in a central force.

of constant $k'$. Let the (horizontal) surface on which these objects slide be frictionless, and let $x_1(t)$ and $x_2(t)$ be the displacement of each object from its equilibrium point. Then, Newton’s second law yields

$$m \frac{d^2x_1}{dt^2} = -kx_1 + k'(x_2 - x_1) \quad ; \quad m \frac{d^2x_2}{dt^2} = -kx_2 - k'(x_2 - x_1)$$

(11.35)

for the equations of motion. To reduce the number of parameters, however, we recast these equations in dimensionless form by selecting a unit of length $a$, dividing the equations by $a$, setting $x_i/a = \xi_i$, introducing a dimensionless time variable $\tilde{t} = \omega t$, where $\omega = \sqrt{k/m}$, and setting $\kappa = k'/k$. The equations then become

$$\frac{d^2\xi_1}{d\tilde{t}^2} = -\xi_1 + \kappa (\xi_2 - \xi_1) \quad ; \quad \frac{d^2\xi_2}{d\tilde{t}^2} = -\xi_2 - \kappa (\xi_2 - \xi_1)$$

(11.36)

Initial conditions such as

$$\xi_1(0) = \xi_{10} \quad ; \quad \frac{d\xi_1}{dt}(0) = \upsilon_{10} \quad ; \quad \xi_2(0) = \xi_{20} \quad ; \quad \frac{d\xi_2}{dt}(0) = \upsilon_{20}$$

(11.37)

complete the statement of the problem—a problem that involves a pair of coupled, linear, second-order, constant coefficient, homogeneous differential equations containing one internal parameter $\kappa$.

11.1.7 Motion under Central Forces

Consider next an object of mass $m$ moving in the $xy$ plane under the action of a central force$^8 \mathbf{F}$, as shown in Fig. 11.6. According to Newton’s second law, the position vector $\mathbf{r}$ of this object satisfies

$^8$A central force is one whose direction is always away from or towards a fixed point—the force center—and whose magnitude depends only on the distance from that point.
the differential equation
\[ m\frac{d^2\mathbf{r}}{dt^2} = \mathbf{F} = f(r) \hat{r} \] (11.38)

where \( f(r) > 0 \) corresponds to a repulsive force and \( f(r) < 0 \) corresponds to an attractive force. Two articulations of this equation are in order. A more specific expression in polar coordinates is treated in an exercise. Here, we extract its components in Cartesian coordinates \((x,y)\). Since
\[ r^2 = x^2 + y^2, \quad r = x \hat{i} + y \hat{j}, \quad \text{and} \]
\[ \hat{r} = \frac{r}{r} = \frac{x \hat{i} + y \hat{j}}{\sqrt{x^2 + y^2}} \] (11.39)
the Cartesian components of Eq. (11.38) are
\[ m\frac{d^2x}{dt^2} = f\left(\sqrt{x^2 + y^2}\right) \frac{x}{\sqrt{x^2 + y^2}} ; \quad m\frac{d^2y}{dt^2} = f\left(\sqrt{x^2 + y^2}\right) \frac{y}{\sqrt{x^2 + y^2}} \] (11.40)

We have arrived at a pair of second-order, non-linear, coupled differential equations to be solved subject to general initial conditions of the form
\[ x(0) = x_0 ; \quad \frac{dx}{dt}(0) = v_x(0) = v_{x0} ; \quad y(0) = y_0 ; \quad \frac{dy}{dt}(0) = v_y(0) = v_{y0} \] (11.41)

Everything discussed so far in this subsection applies to all central forces regardless of the specific form or sign of \( f(r) \). To set a more specific problem, let us narrow our purview to the planetary problem, in which a planet of mass \( m \) orbits a central sun of mass \( M \). When \( M \gg m \), as is often the case, the sun does not move appreciably under the action of the gravitational force exerted on it by the planet. Thus, we can treat the gravitational force on the planet as originating in a fixed force center, in which case
\[ f(r) = -G\frac{mM}{r^2} = -G\frac{mM}{x^2 + y^2} \] (11.42)

where \( G \) is the universal gravitational constant. With this specific force, the members of Eq. (11.40) become
\[ \frac{d^2x}{dt^2} = -GM \frac{x}{(x^2 + y^2)^{3/2}} ; \quad \frac{d^2y}{dt^2} = -GM \frac{y}{(x^2 + y^2)^{3/2}} \] (11.43)

Particularly in the context of this problem, casting the fundamental equations in dimensionless form is prudent. We begin by choosing a reference length, symbolized by \( \ell \). Then, we express all distances in the equations as multiples of this chosen reference by introducing the variables
\[ \mathcal{x} = \frac{x}{\ell} ; \quad \mathcal{y} = \frac{y}{\ell} ; \quad \mathcal{r} = \frac{r}{\ell} \] (11.44)

With this change, Eq. (11.43) becomes
\[ \frac{d^2\mathcal{x}}{d\mathcal{t}^2} = -GM \frac{\mathcal{x}}{(\mathcal{x}^2 + \mathcal{y}^2)^{3/2}} ; \quad \frac{d^2\mathcal{y}}{d\mathcal{t}^2} = -GM \frac{\mathcal{y}}{(\mathcal{x}^2 + \mathcal{y}^2)^{3/2}} \] (11.45)

Finally, we introduce the dimensionless time variable \( \mathcal{t} = \mathcal{t} \sqrt{GM/\ell^3} \) to find that
\[ \frac{d^2\mathcal{x}}{d\mathcal{t}^2} = -\frac{\mathcal{x}}{(\mathcal{x}^2 + \mathcal{y}^2)^{3/2}} ; \quad \frac{d^2\mathcal{y}}{d\mathcal{t}^2} = -\frac{\mathcal{y}}{(\mathcal{x}^2 + \mathcal{y}^2)^{3/2}} \] (11.46)

Interestingly, in Cartesian coordinates, all parameters have disappeared from the equations.

---

9Remember, too, that the problem of a charged particle moving in the field of another charged particle is mathematically identical to the planetary problem. (See exercises.)
Translation of the initial conditions into dimensionless form is easier than translation of the differential equations. The positions, of course, become

$$x(0) = \frac{x_0}{\ell}; \quad y(0) = \frac{y_0}{\ell}$$

To translate the velocities, we argue that

$$\frac{dx}{dt}(0) = \frac{d(x/\ell)}{d(t\sqrt{GM/\ell^3})}(0) = \frac{v_{x0}}{\sqrt{GM/\ell}}; \quad \frac{dy}{dt}(0) = \frac{d(y/\ell)}{d(t\sqrt{GM/\ell^3})}(0) = \frac{v_{y0}}{\sqrt{GM/\ell}}$$

We conclude that dimensionless velocities in the present context are measured in units of $\sqrt{GM/\ell}$.

Before leaving this important problem, we note two additional features. First, if the planet happens to be moving at a distance $r_{\text{circ}}$ from the sun with a velocity of magnitude $v_{\text{circ}}$ directed perpendicular to the radius line, then the planet will move in a circular orbit if its speed and radius are related so that the centripetal force $mv_{\text{circ}}^2/r_{\text{circ}}$ needed for circular motion is exactly provided by the gravitational attraction $GMm/r_{\text{circ}}^2$ of the sun. That is, the orbit will be circular if

$$\frac{mv_{\text{circ}}^2}{r_{\text{circ}}} = \frac{GMm}{r_{\text{circ}}^2} \implies v_{\text{circ}}^2 = \frac{GM}{r_{\text{circ}}}$$

Translated into dimensionless form, this special relationship becomes

$$\left(\frac{\sqrt{GM/\ell} v_{\text{circ}}}{r_{\text{circ}}} \right)^2 = \frac{GM}{\ell r_{\text{circ}}} \implies v_{\text{circ}}^2 = \frac{1}{r_{\text{circ}}}$$

In the dimensionless units we have chosen, the relationship between speed and radius for a circular orbit involves no dimensional constants. This simple case provides us with a specific known motion against which we can later test numerical solutions for the planetary problem.

Second, the planetary problem admits two conservation laws, each of which may be valuable in assessing the accuracy of numerically generated solutions. In a dimensional presentation, conservation of energy yields that

$$E = \frac{1}{2} m \left( v_x^2 + v_y^2 \right) - \frac{GMm}{\sqrt{x^2 + y^2}} = \text{constant}$$

though the mass $m$ can be omitted from the expression if only the constancy of $E$ (and not its actual value) is to be examined. Similarly, in a dimensional presentation, conservation of angular momentum yields that

$$L = m \left( xv_y - yv_x \right) = \text{constant}$$

where, again, $m$ can be omitted if only the constancy of $L$ is to be assessed. Recasting these expressions in dimensionless form (and omitting overall multiplying constants), we find alternatively that

$$\frac{1}{2} \left( \overline{v_x}^2 + \overline{v_y}^2 \right) - \frac{1}{\sqrt{\overline{x}^2 + \overline{y}^2}} = \text{constant}; \quad \overline{x} \overline{v}_y - \overline{y} \overline{v}_x = \text{constant}$$

### 11.1.8 Standing Waves in a String

When a wave propagates in a flexible string, the (transverse) displacement $u(x,t)$ at time $t$ of the element of the string nominally at coordinate $x$ satisfies the wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}$$
where \( c \) is the speed of propagation of the wave along the string. In the special case that the motion of each element is sinusoidal with frequency \( \omega \), we can write

\[
u(x, t) = f(x) \cos(\omega t)\]

and find, on substitution into the wave equation, that \( f(x) \), which gives the amplitude of the sinusoidal motion of the element nominally at \( x \), satisfies

\[
\frac{d^2f}{dx^2} + k^2f = 0
\]

where \( k^2 = \frac{\omega^2}{c^2} \). If, finally, the string is firmly tied down at two points, say \( x = 0 \) and \( x = \ell \), then this second-order, homogeneous, constant-coefficient, differential equation must be solved subject to the boundary conditions

\[
f(0) = 0 \quad ; \quad f(\ell) = 0
\]

and we conclude that waves in a string fixed at two points are described by a boundary value problem. Ultimately, we shall be able to find acceptable solutions only for a discrete set of special values of \( k \).

### 11.1.9 The Schrödinger Equation in One Dimension

A quantum mechanical particle of mass \( m \) having definite energy \( E \) and confined in one dimension \( x \) by a potential energy \( V(x) \) is described by a wave function \( \psi(x) \) that satisfies the one-dimensional time-independent Schrödinger equation

\[
\frac{-\hbar^2}{2m} \frac{d^2\psi}{dx^2} + V(x) \psi = E \psi \quad \Rightarrow \quad \frac{d^2\psi}{dx^2} + \frac{2m}{\hbar^2} \left( E - V(x) \right) \psi = 0
\]

where \( \hbar \) is Planck’s constant divided by \( 2\pi \). We anticipate subsequent exploration of two specific cases. Suppose, for example, that the potential energy is infinite except in the interval \( -\ell \leq x \leq +\ell \), in which interval it is zero, i.e., suppose the particle is confined in an infinitely deep potential well that extends over the specified interval. Then, the wave function describing this particle must satisfy the Schrödinger equation with \( V = 0 \) inside this interval, and it must go to zero at each end of the interval. We seek solutions to the boundary value problem

\[
\frac{d^2\psi}{dx^2} + \frac{2mE}{\hbar^2} \psi = 0 \quad ; \quad \psi(0) = \psi(\ell) = 0
\]

Especially if we substitute the shorthand \( k^2 = \frac{2mE}{\hbar^2} \) (or \( E = \frac{\hbar^2 k^2}{2m} \)), we recognize that this quantum mechanical problem is mathematically identical to the classical standing wave problem described in Section 11.1.8. This problem admits solutions only for a discrete set of special values for \( k \), i.e., only for special energies.

Suppose, alternatively, that we take the potential energy to be \( V(x) = \frac{1}{2}kx^2 \), which describes the quantum analog to the classical harmonic oscillator with spring constant \( k \). In this case, the time-independent Schrödinger equation becomes

\[
\frac{d^2\psi}{dx^2} + \frac{2m}{\hbar^2} \left( E - \frac{1}{2}kx^2 \right) \psi = 0
\]

or, in dimensionless form,

\[
\frac{d^2\psi}{d\tau^2} + (2\epsilon - \tau^2) \psi = 0
\]

where \( \omega = \sqrt{k/m} \), \( \epsilon = E/\hbar\omega \) and \( \tau = x\sqrt{\hbar/m\omega} \). In contrast to the infinitely deep well, the domain for the quantum oscillator extends over the interval \( -\infty < \tau < +\infty \), and we must require that \( \psi(x) \) approach zero as \( x \) approaches either end of this interval. Numerically, infinite domains
are complicated. Note, however, that the equation in this case is quadratic in \( x \). Thus, the solutions to the equation can be divided into two sets, one of which contains functions that are even in \( x \) and the other of which contains functions that are odd in \( x \). Even functions, however, necessarily have zero derivative and non-zero value at \( x = 0 \) while odd functions, in contrast, have non-zero derivative and zero value at \( x = 0 \). These properties mean that we can replace the original boundary value problem involving Eq. (11.60) over an infinite interval with an initial value problem for the solution over half of the infinite interval. When we come later to address the actual solution of this problem, we will therefore focus on two sub-problems, one defined by the equation and initial values

\[
\frac{d^2 \psi}{dx^2} + (2\epsilon - \pi^2) \psi = 0 ; \quad \psi(0) = 1.0 ; \quad \frac{d\psi}{dx}(0) = 0 ; \quad \lim_{\pi \to \infty} \psi(\pi) = 0
\] (11.62)

and yielding even solutions and the other defined by the equation and initial values

\[
\frac{d^2 \psi}{dx^2} + (2\epsilon - \pi^2) \psi = 0 ; \quad \psi(0) = 0 ; \quad \frac{d\psi}{dx}(0) = 1 ; \quad \lim_{\pi \to \infty} \psi(\pi) = 0
\] (11.63)

and yielding odd solutions. The interesting aspect of these problems is that we will be able to find solutions satisfying the boundary requirement at \( \pi = +\infty \) only for very special values of \( \epsilon \) in the equation, i.e., only for special values of the energy.

11.2 Laplace Transforms

One tool used behind the scenes by symbolic solvers of ODEs is called the Laplace transform, which we defined and illustrated in Section 1.5.2. We return to the topic here, not because we are likely to make much use of it directly but because knowing its properties and capabilities may sometimes be valuable as we try to guide a symbolic manipulator that uses the technique. As we laid out in the referenced section, the Laplace transform is defined for a function \( f(t) \) by the integral

\[
\mathcal{L}(f(t)) = \tilde{f}(s) = \int_0^\infty e^{-st} f(t) \, dt
\] (11.64)

We can best illustrate the value of the Laplace transform in solving a linear, ordinary differential equation with constant coefficients by presenting a short example. We would—and symbolic manipulators may well—apply this technique to the third member of Eq. (11.4) as follows. First we use Eqs. (1.16) and (1.17) and entries in Table 1.2 to evaluate its Laplace transform, finding that

\[
m[s^2 \tilde{z}(s) - sz_0 - v_{z0}] = -\frac{mg}{s} - b[s \tilde{z}(s) - z_0]
\] (11.65)

Then we solve this algebraic equation\(^\text{12}\) for \( \tilde{z}(s) \), finding that

\[
\tilde{z}(s) = \frac{z_0}{s} + \frac{v_{z0} s - g}{s^2 (s + b/m)}
\] (11.66)

Finally, we invert this Laplace transform to find the solution itself. Unfortunately, inverting Laplace transforms is rarely easy.\(^\text{13}\) We can, of course, read a table of transforms such as Table 1.2 backwards,

\(^{10}\) It involves only second derivatives and the potential energy involves only \( x^2 \).

\(^{11}\) Reducing infinity to half of infinity may not seem like much of a simplification. The primary value of the change is that the starting point for tracking a solution has now been moved into an accessible region.

\(^{12}\) The immense simplification coming from using the Laplace transform is precisely the conversion of a differential equation (of the right sort) into an algebraic equation for the transform of the solution.

\(^{13}\) Murphy’s law, in the version that says that transforms may move difficulties around but may not eliminate them altogether, applies. The price we pay for the conversion to an algebraic equation is that we now have to invert the transform we found so easily.
but that strategy will work only if the entry we need is in fact included in the table. In the present case, however, we can recast the result we have obtained by using the technique of partial fractions to find the equivalent expression

\[
\tilde{z}(s) = \frac{z_0}{s} - \frac{mg}{bs} + \frac{m}{b} \left( v_{z0} + \frac{mg}{b} \right) \left( \frac{1}{s} - \frac{1}{s + b/m} \right) \tag{11.67}
\]

All of the pieces in this form of the transform now are in even the tiny table we have available. Since the inverse of a sum of terms is the sum of the inverses of each term separately, we conclude that

\[
z(t) = z_0 - \frac{mg}{b} t + \frac{m}{b} \left( v_{z0} + \frac{mg}{b} \right) \left( 1 - e^{-bt/m} \right) \tag{11.68}
\]

Note that, in this approach, the initial conditions imposed on the solution are incorporated ab initio, not imposed after a general solution with arbitrary, unknown constants has been obtained.

### 11.5 Solving ODEs Symbolically with Mathematica

Fundamentally, solving an ordinary differential equation involves specifying the equation, finding a solution containing the appropriate number of arbitrary constants, and imposing the appropriate initial or boundary conditions. A few features of the command DSolve, which is the primary ODE-solving command in Mathematica, were discussed briefly in Section 8.8.9. In that section, we used DSolve only for solving a single ODE, and we learned that—to that end—the command has two forms, each of which itself has two (sub)forms. To use the first form, we submit either the statement

\[
\text{DSolve[ ODE, DVar[ IVar], IVar ]}
\]

or the statement

\[
\text{DSolve[ ODE, DVar, IVar ]}
\]

In either case, the command returns the solution of the single ordinary differential equation ODE for the dependent variable DVar as a function of the single independent variable IVar. In the first subform, the solution is presented as a transformation rule for \( x[t] \); in the second subform, Mathematica instead returns a rule defining \( x \) as a function of the variable \( t \). In the form of this paragraph, DSolve does not impose initial conditions; they haven’t even been specified. Instead, it uses the symbols \( C[1], C[2], \ldots \) (however many are needed) for the integration constants. These constants must then be determined by subsequent imposition of appropriate initial conditions.

In the second form, the command DSolve accepts as its first argument a list which contains both the differential equation to be solved and an appropriate number of initial conditions. The statement would then have either the first subform

\[
\text{DSolve[ \{ ODE, IC1, IC2, \ldots \}, DVar[ IVar], IVar ]}
\]

or the second subform

\[
\text{DSolve[ \{ ODE, IC1, IC2, \ldots \}, DVar, IVar ]}
\]

where these statements return results with the same features as described in the previous paragraph for the first form. These forms return solutions on which the initial conditions IC1, IC2, \ldots have been imposed and which therefore contain no undetermined constants. If, for example, the function to be found is \( y[t] \), then an initial condition on the function itself would have the form \( y[0] == y_0 \) while an initial condition on its first derivative would have the form \( y'[0] == v_0 \), where \( y_0 \) and \( v_0 \)
are the desired values—and may be symbolic or explicitly numeric. Note, in particular, the use of the symbol \( \equiv \) to specify those initial conditions.

The command `DSolve` is also capable of solving (some) systems of ordinary differential equations. In this case, the first argument will be a list containing all of the differential equations as well as any initial conditions to be imposed, the second argument will be a list containing the several dependent variables, and the third argument will identify the independent variable. In brief, the statement invoking `DSolve` will have the subform

\[
\text{DSolve[ \{ ODE1, ODE2, ..., IC1, IC2, ...\}, \{ DVar1[IVar], DVar2[IVar], ...\}, IVar ]}
\]

or, alternatively, the subform

\[
\text{DSolve[ \{ ODE1, ODE2, ..., IC1, IC2, ...\}, \{ DVar1, DVar2, ...\}, IVar ]}
\]

### 11.5.1 Projectile in a Viscous Medium

For the sake of a comparison with the result obtained in Section 11.2, we begin by using `Mathematica` to solve the third member of Eq. (11.4). The necessary commands are\(^\text{14}\)

\[
\begin{align*}
\text{In[1]:=} \quad & \text{eq = m*z''[t] == -m*g - b*z'[t]} \\
\text{Out[1]=} \quad & m z''[t] == -gm - bz'[t] \\
\text{In[2]:=} \quad & \text{soln = DSolve[ eq, z, t ][[1,1]]} \\
\text{Out[2]=} \quad & z \rightarrow \text{Function}\left[\left\{t, \frac{gm t}{b} - \frac{e^{-bt/m} m C[1]}{b} + C[2]\right\}\right]
\end{align*}
\]

Note that `Mathematica` has here tacitly assumed that \( b \neq 0 \). More careful treatment would be required if \( b = 0 \), i.e., in the case of no damping.

To complete the solution, we must of course impose appropriate initial conditions. Thus, we continue with the statements

\[
\begin{align*}
\text{In[3]:=} \quad & \text{eq1 = z[0] == z0 /. soln;} \\
\text{In[4]:=} \quad & \text{eq2 = z'[0] == vz0 /. soln;} \\
\text{In[5]:=} \quad & \text{Solve[ \{eq1,eq2\}, \{C[1],C[2]\} ][[1]]}$ \\
\text{In[6]:=} \quad & \text{soln = soln /. %;} \\
\text{In[7]:=} \quad & \text{z[t] /. soln;} \\
\text{In[8]:=} \quad & \text{z'[t] /. soln;}
\end{align*}
\]

The end result of these operations is the solution

\[
z(t) = z_0 - \frac{mg}{b} t + \frac{m}{b} \left(v_{z0} + \frac{mg}{b}\right) \left(1 - e^{-bt/m}\right) \tag{11.69}
\]

\(^\text{14}\)As in Chapter 8, we abbreviate the presentation of `Mathematica` dialogs by making liberal use of terminating semicolons to suppress intermediate output. You are urged to duplicate the dialog in an actual session with `Mathematica`, omitting all terminating semicolons.
for the position and the solution
\[ v_z(t) = -\frac{mg}{b} + \left( v_{z0} + \frac{mg}{b} \right) e^{-bt/m} \]
(11.70)
for the velocity, though we have rearranged the solutions presented by Mathematica to improve their readability. This result for \( z(t) \) agrees fully with the result obtained in Eq. (11.68).

Wisdom suggests that we should whenever possible actually verify any solution that Mathematica generates. To this end, we substitute the solution into the original equation, evaluate the derivatives explicitly, and simplify the result with the statement

\[
\text{In}[9]:= \text{Simplify[ eq /. soln ]}
\]
\text{Out}[9]= True

To complete the verification, we check satisfaction of the initial conditions with the statement

\[
\text{In}[10]:= \text{Simplify[ \{ z[0], z'[0] \} /. soln ]}
\]
\text{Out}[10]= \{ \text{z0}, \text{vz0} \}

All is indeed well.

One further check on this solution involves examining its limit as \( b \)—the viscous damping—becomes small. For the coordinate and velocity, we find respectively that

\[
\text{In}[11]:= \text{Collect[ Series[ z[t] /. soln, \{ b, 0, 1 \}], t ]}
\]
\text{Out}[11]= \text{z0} + v_{z0}t - \frac{1}{2}gt^2 - \frac{bt^2}{2m} \left( v_{z0} - \frac{1}{3}gt \right) + O[b]^2
\]
\[
\text{In}[12]:= \text{Collect[ Series[ z'[t] /. soln, \{ b, 0, 1 \}], t ]}
\]
\text{Out}[12]= v_{z0} - gt - \frac{bt}{m} \left( v_{z0} - \frac{1}{2}gt \right) + O[b]^2
\]
The terms free of \( b \) (i.e., terms in \( b^0 = 1 \)) clearly agree with the known results for free fall in the absence of viscous resistance.

The solution of this problem when motion occurs in two or three of the coordinate directions is explored in one of the exercises.

We have, of course, taken a bit longer route than necessary so as to illustrate the use of Mathematica to find the undetermined constants in the general solution to an ODE. We could combine the first several of our statements into an invocation of DSolve that includes the incorporation of the initial conditions with the statements

\[
\text{In}[13]:= \text{Quit[]}
\]
\[
\text{In}[1]:= \text{eq = m*z''[t] == -m*g-b*z'[t];}
\]
\[
\text{In}[2]:= \text{soln = DSolve[ \{eq, z[0]==z0, z'[0]==vz0\}, z, t ];}
\]
\text{In}[3]:= \text{Quit[]}
\]
These statements lead us in rather fewer steps to the result already presented in Eq. (11.69), though it may require extraction of the expression for the solution from soln and a subsequent application of Expand and Collect to reveal that agreement more clearly.

\footnote{Again, we have taken the liberty to rearrange the terms in Mathematica’s output to facilitate recognition of the limits.}
11.5.2 Logistic Growth

Consider next Eq. (11.22) governing the logistic growth of a population \( N \) in an environment with carrying capacity \( N_c \). We seek the solution with the Mathematica statements

\[
\text{In}[1]:= \text{eq} = n'[t] == k*n[t]*(1-n[t]/n_c); \\
\text{In}[2]:= \text{soln} = \text{DSolve}[\{\text{eq}, n[0]==n0\}, n, t ];
\]

Solve::ifun : Inverse functions are being used by Solve, so some solutions may not be round.

\[
\text{Out}[2]= \left\{ \text{n} \rightarrow \text{Function}\left[ \{t\}, \frac{\text{e}^{k t n_0 n_c}}{-n_0 + \text{e}^{k t n_0 + n_c}} \right] \right\}
\]

Enter equation. (We use \( n \) rather than \( N \) because \( N \) is a reserved symbol in Mathematica.)

Solve equation for \( n \) as function of \( t \), imposing the general initial condition \( n(0) = n_0 \). Mathematica has produced a warning message that, in this context, we can safely ignore.

Elevate function definition to top level.

We can verify satisfaction of the initial condition and the differential equation with the statements

\[
\text{In}[4]:= n[0] /. \text{soln} \\
\text{Out}[4]= n_0 \\
\text{In}[5]:= \text{Simplify}[\text{eq} /. \text{soln}] \\
\text{Out}[5]= \text{True}
\]

Verify satisfaction of initial condition. Verify satisfaction of ODE.

Verifying the limit as \( t \to \infty \) is more complicated because, in the most straightforward statement, Mathematica's command Limit will assume \( k \) to be complex and the command has no feature allowing us to tell Mathematica otherwise. Instead, we use the statements

\[
\text{In}[6]:= \text{qn} = n[t] /. \text{soln} \\
\text{Out}[6]= \frac{\text{e}^{k t n_0 n_c}}{-n_0 + \text{e}^{k t n_0 + n_c}} \\
\text{In}[7]:= \text{qn} /. \text{Exp}[k*t] \to \text{tmp} \\
\text{Out}[7]= \frac{\text{tmp} n_0 n_c}{-n_0 + \text{tmp} n_0 + n_c}
\]

to obtain an expression giving the solution and then cast that expression in terms of the single variable \( \text{tmp} \) for \( e^{kt} \). Then, since \( k > 0 \) and—hence—\( e^{kt} = \text{tmp} \to \infty \) as \( t \to \infty \), we can determine the solution at large \( t \) with the statement

\[
\text{In}[8]:= \text{Limit}[\%, \text{tmp} \to \text{Infinity}] \\
\text{Out}[8]= n_c
\]

As expected, the solution approaches the carrying capacity, regardless of the initial population. The solution therefore passes all three tests.

Let us conclude the discussion of this example by displaying the evolution of the population in a graph. We begin by casting the expression to be plotted in a dimensionless form, choosing \( n_c \) as the unit in terms of which to express the population. We have in line \( \text{Out}[6] \) already obtained the expression to be plotted. We proceed with the statements

\[
\text{In}[9]:= \text{qn} = \text{qn}/\text{nc} \\
\text{Out}[9]= \frac{\text{e}^{k t n_0}}{-n_0 + \text{e}^{k t n_0 + n_c}}
\]

Measure population \( n \) in units of \( n_c \).
202

CHAPTER 11. SOLVING ORDINARY DIFFERENTIAL EQUATIONS

In[10]:= qn=Simplify[ qn/.n0->\[Alpha]*nc ]

Out[10]=

\[ e^{kt} \alpha \]

Finally, we make one further recasting to measure time in units of \(1/k\) with the statement

In[11]:= qn = qn /. t -> \[Tau]/k

Out[11]=

\[ e^{\tau} \alpha \]

With these rescalings, we have but one parameter [physically (biologically?) the initial population measured in units of the carrying capacity]. We produce the graph in Fig. 11.7 with the statements

In[12]:= q1 = qn /. \[Alpha] -> 0.25;
In[13]:= q2 = qn /. \[Alpha] -> 0.5;
In[14]:= q3 = qn /. \[Alpha] -> 1.0;
In[15]:= q4 = qn /. \[Alpha] -> 2.0;
In[16]:= q5 = qn /. \[Alpha] -> 4.0;
In[17]:= Plot[ {q1,q2,q3,q4,q5}, \{[Tau], 0.0, 5.0\}, PlotRange -> \{0.0, 4.0\},
PlotStyle -> Thickness[0.01], PlotLabel->StyleForm[ "Logistic Behavior",
FontFamily->"Times", FontSize -> 14,
FontWeight->"Bold"], AxesLabel->\{"kt", "n/nc"\} ]

Regardless of the initial population, all solutions converge monotonically on the carrying capacity (1.0 in the units we are using). Further, if we execute the statements

In[18]:= q6 = qn /. \[Alpha] -> 0.05;
In[19]:= Plot[ q6, \\{[Tau], 0.0, 10.0\},
PlotRange -> \{0.0, 1.0\}, PlotLabel->StyleForm[ "Logistic Behavior",
FontFamily->"Times", FontSize -> 14,
FontWeight->"Bold"], AxesLabel->\{"kt", "n/nc"\} ]

we generate the graph in Fig. 11.8, which shows the initial exponential growth when the population is well below the carrying capacity but then reveals the leveling off as the population approaches the carrying capacity.

11.5.3 Damped Harmonic Oscillator

A third example of the use of Mathematica to solve a single second-order, linear, constant-coefficient, homogeneous, ordinary differential equation was presented in Chapter 8. You are invited to review the discussion of the damped harmonic oscillator in Sections 8.8.9 and 11.1.4.

11.5.4 Chain Radioactive Decay

As a first example of the use of Mathematica to solve a system of linear equations, let us determine the behavior of the radioactive decay chain described by Eq. (11.19). For simplicity, we suppose the initial values \(A(0) = A_0\) and \(B(0) = C(0) = 0\), i.e., we start with some \(A\) and no \(B\) or \(C\). The temporal evolution of this system is found with the Mathematica statements
Figure 11.7: Logistic growth or decay of a population when, starting with the highest graph, the initial population is 4.0, 2.0, 1.0, 0.5, and 0.25 times the carrying capacity of the environment.

![Logistic Behavior](image1)

Figure 11.8: Logistic growth when the population is initially much smaller than the carrying capacity. This curve is called the *sigmoid* curve.

![Logistic Behavior](image2)

\[
\begin{align*}
\text{In}[1]:&= \text{eqA} = A'[t] == -kA*A[t]; \\
\text{In}[2]:&= \text{eqB} = B'[t] == kA*A[t] - kB*B[t]; \\
\text{In}[3]:&= \text{eqC} = C'[t] == kB*B[t]; \\
\text{In}[4]:&= \text{soln} = \text{DSolve}\{\text{eqA, eqB, eqC, A[0]==A0, B[0]==0, C[0]==0, } \{A,B,C, t \}
\end{align*}
\]

Enter equations, suppressing display.

Request solution, incorporating initial conditions in request.

*Mathematica* displays the fairly complicated result
To cast these results in a more interpretable form, let us extract each solution as an expression with the statements

\[ \text{In}[5] := qA = A[t] \/. \text{soln[[1]]} \]
\[ \text{Out}[5] = A_0 e^{-k_A t} \]

\[ \text{In}[6] := qB = \text{Simplify}[B[t] \/. \text{soln[[1]]}] \]
\[ \text{Out}[6] = A_0 \left( e^{-k_A t} - e^{-k_B t}\right) \frac{k_A}{k_A - k_B} \]

\[ \text{In}[7] := qC = \text{Simplify}[C[t] \/. \text{soln[[1]]}] \]
\[ \text{Out}[7] = A_0 \left( k_A e^{-k_A t} + (1 + e^{-k_A t}) k_B \right) \frac{k_B}{k_A - k_B} \]

These forms make evident one important conclusion: in this chain decay, we see a linear combination of two exponential decays, each with its own distinct half-life.

We conclude the discussion of this example by generating a graph of \( A(t) \), \( B(t) \), and \( C(t) \) for a specific set of values. Anticipating that we will set \( k_A = k_B \) for the graphs and noting that the expressions obtained for \( B(t) \) and \( C(t) \) are indeterminate when \( k_A = k_B \), we prepare for the graph by evaluating the limits of \( qB \) and \( qC \) as \( k_B \to k_A \). The statements

\[ \text{In}[8] := qBeq = \text{Limit}[qB, kB \to kA] \]
\[ \text{Out}[8] = A_0 e^{-k_A t} \]

\[ \text{In}[9] := qCeq = \text{Limit}[qC, kB \to kA] \]
\[ \text{Out}[9] = A_0 \left( k_A e^{-k_A t} + (1 + e^{-k_A t}) k_B \right) \frac{k_B}{k_A - k_B} \]

will evaluate those limits, leaving each expressed in terms of the parameters \( k_A \) and \( A_0 \). Then, to obtain the desired graph, we specify values for \( A_0 \) and \( k_A \) with the statements

\[ \text{In}[10] := A0 = 1000; kA = 0.1; \]

and produce the graph with the single statement

\[ \text{In}[11] := \text{Plot}[\{qA, qBeq, qCeq\}, \{t, 0.0, 50.0\}, \text{PlotRange} \to \{0.0, 1000.0\}, \text{PlotStyle} \to \text{Thickness}[0.005], \text{PlotLabel} \to \text{StyleForm["Radioactive Decay", FontFamily->"Times", FontSize->14, FontWeight->"Bold"], AxesLabel} \to \{"t", "A,B,C\} ] \]

The resulting graph is shown in Fig. 11.9.
11.5. SOLVING ODES SYMBOLICALLY WITH MATHEMATICA

Figure 11.9: Radioactive decay of A, B, and C. The graph starting at 1000 is A; the graph rising to 1000 at $t = 50.0$ is C; and the remaining graph is B.

11.5.5 Coupled Oscillators

Let us next solve for the motion of the coupled oscillators described in Section 11.1.6. We shall address the problem in the dimensionless form presented in Eq. (11.36), imposing the general initial conditions in Eq. (11.37). For simplicity, however, we will take the initial velocities both to be zero.\(^\text{16}\) First, suppressing the display of Mathematica’s response, we enter the equations with the statements:\(^\text{17}\)

\[
\text{In}[1]:= \text{eq1} = x_1''[t] + x_1[t] - \kappa (x_2[t] - x_1[t]) = 0; \\
\text{In}[2]:= \text{eq2} = x_2''[t] + x_2[t] + \kappa (x_2[t] - x_1[t]) = 0;
\]

Then we create a list containing the initial values with the statement

\[
\text{In}[3]:= \text{ics} = \{ x_1[0]==x_{10}, \ x_2[0]==x_{20}, \ x_1'[0]==0, \ x_2'[0]==0 \};
\]

and, finally, we request the solution with the statement

\[
\text{In}[4]:= \text{soln} = \text{DSolve}[\text{Union}\{\text{eq1,eq2}, \text{ics}\}, \{x_1, \ x_2\}, \text{t}]; \\
\text{Out}[4]= \{\{x_1 \rightarrow \text{mess}, \ x_2 \rightarrow \text{mess}\}\}
\]

While correct, the (suppressed) output from this statement is unfortunately in a form that is difficult to fathom. To rectify that situation, we

\begin{itemize}
  \item execute the statement
    
    \[
    \text{In}[5]:= \text{soln1} = \text{soln} /. \text{Sqrt}[-1-2*\kappa] \rightarrow \text{I*sqrt}[1+2*\kappa];
    \]
\end{itemize}

\(^{16}\)You are urged to explore other initial conditions.

\(^{17}\)In the derivation of the dimensionless equations, we used overbars to identify the dimensionless quantities. We here drop that refinement in the notation, omitting the overbars but understanding that the quantities represented are still dimensionless.
to replace several occurrences of $\sqrt{-1 - 2\kappa}$ with $i\sqrt{1 + 2\kappa}$, thus making the imaginary character of $\sqrt{-1 - 2\kappa}$ explicit. (Remember that $\kappa > 0$.)

- extract expressions for the position of each oscillator as a function of time with the statements

In[6] := qx1 = x1[t] /. soln1[[1]];  
In[7] := qx2 = x2[t] /. soln1[[1]];  

- execute the statements

In[8] := qx1a = Expand[ qx1 ];  
In[9] := qx2a = Expand[ qx2 ];

to simplify the resulting expressions by multiplying several exponentials together.

- convert complex exponentials to trigonometric functions with the statements

In[10] := qx1b = ExpToTrig[ qx1a ];  
In[11] := qx2b = ExpToTrig[ qx2a ];

At this point, we recognize that the time dependence of each term in these solutions has either $\cos[t]$ or $\cos[t\sqrt{1 + 2\kappa}]$ as a factor. Thus, we collect those terms with the nested statements

In[12] := qx1c = Collect[ Collect[qx1b, Cos[t]], Cos[t*Sqrt[1 + 2*Kappa]] ]  
Out[12] = \( \left( \frac{x_{10}}{2} + \frac{x_{20}}{2} \right) \cos[t] + \left( \frac{x_{10}}{2} - \frac{x_{20}}{2} \right) \cos[t\sqrt{1 + 2\kappa}] \)

In[12] := qx2c = Collect[ Collect[qx2b, Cos[t]], Cos[t*Sqrt[1 + 2*Kappa]] ]  
Out[12] = \( \left( \frac{x_{10}}{2} + \frac{x_{20}}{2} \right) \cos[t] + \left( -\frac{x_{10}}{2} + \frac{x_{20}}{2} \right) \cos[t\sqrt{1 + 2\kappa}] \)

At this point, we finally recognize that the solution is a superposition of different sinusoidal oscillations, one at frequency 1 (in units of $\omega = \sqrt{k/m}$) and the other at frequency $\sqrt{2\kappa + 1}$ (again in units of $\omega$). Clearly also, only the term at frequency 1 is present if $x_{20} = x_{10}$ (the lower-frequency normal mode, in which the two objects oscillate with equal amplitude and in phase), and only the term at frequency $\sqrt{2\kappa + 1}$ is present if $x_{20} = -x_{10}$ (the higher-frequency normal mode, in which the objects oscillate with equal amplitude but out of phase).

11.5.6 Standing Waves in a String

Boundary value problems need to be treated differently from initial value problems. To find the standing waves in a string that satisfy Eqs. (11.56) and (11.57), for example, we might enter and solve the equation subject to zero boundary conditions at each end of the string with the statements

In[1] := eq = f''[x] + k^2 * f[x] == 0;  
In[2] := soln = DSolve[ { eq, f[0]==0, f[l]==0 }, f, x ]

Out[2] = \{ f \to Function \[ \{ x \}, \begin{cases} C[1] \sin[\sqrt{k^2} x] & n \in \text{Integers} \& n \geq 1 \& k^2 = \frac{n^2 \pi^2}{l^2} \& \& \ell > 0 \\ 0 & \text{True} \end{cases} \} \} \}
The version of Mathematica here in use recognizes that only the trivial solution exists unless \( k \) has one of the special values \( k = n\pi/l, n \in \text{integer} \). In the special case, we learn that

\[
f(x) = C_1 \sin \frac{n \pi x}{l}
\]  

(11.71)

where \( C_1 \) is undetermined and \( n \) assumes any of the values 1, 2, 3, \ldots (but not the value 0—because the solution reduces to zero in that case—and not negative values—because they really change only the overall sign of the solution and could be seen simply as an alternative choice of \( C_1 \)). Further, since \( \omega = kc \), we would associate with this solution the frequency

\[
\omega_n = k_n c = \frac{n \pi c}{l} = n \omega_1
\]

(11.72)

where \( \omega_1 \) is the fundamental frequency. We find, in particular, that all allowed frequencies are integer multiples of the fundamental frequency.

Earlier versions of Mathematica concluded that only the trivial solution to the ODE can be made to satisfy both boundary conditions and returned only the solution

\[
\text{Out[2]} = \{ \{ f \to \text{Function}[\{x\}, 0] \} \}
\]

to the statement at \text{In}[2] above. These versions of Mathematica failed to recognize that we can find non-trivial solutions, but only if we are prepared to constrain \( k \) to a limited set of values.

In those versions, we must evidently be more deliberate. To help Mathematica recognize that the problem is a bit more subtle, we simply solve the differential equation without specifying the boundary conditions, thereby generating a solution that includes two initially undetermined integration constants. Then, we impose the boundary conditions. First, we solve the equation with the statement

\[
\text{In[3]} \text{ soln = DSolve[ eq, f, x ][][][1,1]}
\]

Out[3]= \( f \to \text{Function}[\{x\}, C[1] \cos[k x] + C[2] \sin[k x]] \)

elevating the solution with \([1,1]\) to the top level in its expression. Next, we determine the consequences of the boundary conditions with the statements

\[
\text{In[4]} \text{ eq1 = 0 == f[0] /. soln} \quad \text{Impose condition at } x = 0.
\]


\[
\text{In[5]} := \text{eq2 = 0 == f[l] /. soln} \quad \text{Impose condition at } x = l.
\]


At this point, we recognize that we have a pair of equations to be solved simultaneously for \( C[1] \) and \( C[2] \). We also recognize, however, that these equations are homogeneous and will in fact have non-trivial solutions only if the determinant of the coefficient matrix happens to be zero. To complete the problem, then, we must extract that coefficient matrix, set its determinant to zero, and solve the resulting equation to find any values of \( k \) that may admit non-trivial solutions for \( C[1] \) and \( C[2] \). To find the coefficient matrix, we invoke the Mathematica statement\(^{18}\)

\[
\text{In[5]} := \text{tmp = CoefficientArrays[ \{eq1, eq2\}, \{C[1], C[2]\} ]}
\]

Out[5]= \{ \text{SparseArray[ mess ]}, \text{SparseArray[ mess ]} \}

whose output is a bit obscure. Even so, the statement\(^{19}\)

\[\text{(footnotes continued)}\]

\(^{18}\)In some earlier versions of Mathematica, the necessary functionality resided in the older command \text{LinearEquationsToMatrices} in the package \text{LinearAlgebra}. While still available, that entire package is now obsolete and may not be included in future versions of Mathematica.

\(^{19}\)In the output from Mathematica, all branches will be the same size. We have here used different sizes to clarify the structure of that output.
In[6]:= tmp1 = Normal[%]
Out[6]=
\[
\begin{pmatrix}
0 & 0 \\
-\cos(\ell k) & -\sin(\ell k)
\end{pmatrix}
\]

yields a more transparent result. Here, the output is a list of lists, the first of which conveys the inhomogenieties in the equations while the second, itself a list of lists, provides the coefficient matrix (though the minus signs in several places are a bit mysterious). We can extract and display the coefficient matrix with the statements

In[7]:= coefmat = tmp1[[2]];
In[8]:= MatrixForm[ coefmat ]

\[
\begin{pmatrix}
-1 & 0 \\
-\cos(\ell k) & -\sin(\ell k)
\end{pmatrix}
\]

Next, we seek values of \(k\) for which the determinant of this matrix will be zero by invoking the statements

In[9]:= charpol = Det[ coefmat ]
Out[9]= \(\sin(\ell k)\)
In[10]:= rts = Solve[ charpol==0, k ]

\[
\begin{cases}
\kappa \rightarrow \text{ConditionalExpression}\left[\frac{2\pi \kappa}{\ell}, \kappa \in \mathbb{Z}\right], \\
\kappa \rightarrow \text{ConditionalExpression}\left[\frac{\pi + 2\pi \kappa}{\ell}, \kappa \in \mathbb{Z}\right]
\end{cases}
\]

\textit{Mathematica} recognizes that this equation has numerous roots.\(^{20}\) Note that the \(\kappa[1]\) in these statements is \emph{not} the same as the \(\kappa[1]\) at \texttt{Out}[2] and \texttt{In}[4]-\texttt{Out}[5] and in Eq. (11.71). To avoid confusion, let’s simplify the appearance here by replacing \emph{this} \(\kappa[1]\) with \(n\) with the statement

In[11]:= rts1 = rts /. \(\kappa[1] \rightarrow n\)

\[
\begin{cases}
\kappa \rightarrow \text{ConditionalExpression}\left[\frac{2\pi \kappa}{\ell}, \kappa \in \mathbb{Z}\right], \\
\kappa \rightarrow \text{ConditionalExpression}\left[\frac{\pi + 2\pi \kappa}{\ell}, \kappa \in \mathbb{Z}\right]
\end{cases}
\]

In this form, it is easier to see that, with \(n\) an integer, \(k = 2n\pi/\ell\) or \(k = (2n+1)\pi/\ell\), i.e., \(k\) is either an even or an odd multiple of \(\pi/\ell\). More simply, \(k = n\pi/\ell\) for \(n\) any integer. With this recasting, we use the statement

In[12]:= rl = \(k \rightarrow n*\pi/\ell\);

to define a transformation rule for substituting \(n\pi/\ell\) for \(k\) in any expression. With this rule, we quickly find the corresponding values of \(\kappa[1]\) and \(\kappa[2]\) by returning to equations eq1 and eq2 and executing the statements

In[13]:= eq1 /. rl
Out[13]= 0 == \(\kappa[1]\)
In[14]:= Simplify[ eq2 /. rl, \texttt{Element}[n,\mathbb{Z}] ]
Out[14]= 0 == \(\kappa[1]\)

\(^{20}\)An earlier version of \textit{Mathematica} returned the single (correct) root \(k = 0\). The behavior has changed in newer versions of \textit{Mathematica}.\]
11.5. SOLVING ODES SYMBOLICALLY WITH MATHEMATICA

Clearly, we must choose $C[1]$ to be 0—a fact we really knew at the very beginning, but we learn nothing about $C[2]$, which evidently remains arbitrary insofar as the conditions of the problem are concerned.

In the end, we discover that we can find acceptable solutions to the original boundary value problem only when $k$ has one of the values $n\pi/\ell$ and $C[1]$ is set to zero. Within Mathematica, we then find that the solution to the original boundary value problem is given by

$$\text{In}[15]:= \text{stdwv} = f[x] /. \text{sln}$$


$$\text{In}[15]:= \text{stdwv} = \text{stdwv} /. \{r1, C[1] \to 0\}$$

$$\text{Out}[15]= C[2] \frac{n\pi x}{\ell}$$

where $n$ assumes any of the values 1, 2, 3, ... (but not the value 0—because the solution reduces to zero in that case—and not negative values—because they really change only the overall sign of the solution and could be seen simply as an alternative choice of $C[2]$). After this more arduous route, we arrive again at the solution expressed in Eqs. (11.71) and (11.72).

11.5.7 Infinite Depth Quantum Well

As pointed out in Section 11.1.9, the quantum problem of a particle in an infinitely deep, one-dimensional, square potential well is mathematically identical to the problem we have just addressed. Since the allowed values of $k$ have turned out to be $n\pi/\ell$, we then conclude that the allowed energies in the quantum problem would be

$$E_n = \frac{\hbar^2 k^2}{2m} = \frac{n^2 \hbar^2 \pi^2}{2m \ell^2}$$

Further, we would associate the (unnormalized) wave function

$$\psi_n(x) = N_n \sin \frac{n\pi x}{\ell}$$

with the energy $E_n$. In the quantum problem, the allowed energies are not integer multiples of the lowest energy; rather, the allowed energies increase in proportion to the squares of the integers.

11.5.8 Finding Solutions by Series Methods

Particularly when analytic methods appear to be yielding little result, we are sometimes tempted to invoke the method of Frobenius, representing the solution to the ODE of interest as a power series of the form

$$f(x) = \sum_{i=0}^{\infty} a_i x^{i+\gamma}$$

and seeking both an indicial equation determining $\gamma$ and recursion relationships determining $a_i$. In the absence of a simple Mathematica command to facilitate this approach to solving an ODE, we must either write our own package (which is beyond the scope of this book) or step Mathematica systematically through the process. For the sake of concreteness, suppose we seek a solution to the equation

$$\frac{d^2 y}{dx^2} + y = 0$$

We might then define a (truncated) series for the solution with the statement

$$\text{In}[1]:= y[x_] := \text{Sum}[a[i] x^{i+\gamma}(i+\{\text{Gamma}\}), \{i, 0, 8\}]$$
(We could, of course, extend the series as far as we wished. We here truncate it with a few terms to keep the output to manageable size.) Then we would evaluate the right-hand side of the ODE with the statement

\[
\text{In}[2] := y''[x] + y[x] \\
\text{Out}[2] = \text{(expression too long to record)}
\]

Not surprisingly, each term has a factor of \(x^\gamma\). Since we ultimately will set this expression equal to zero, we can divide the expression by \(x^\gamma\), and then—since a premultiplying factor of \(1/x^2\) appears—multiply the entire expression by \(x^2\). The necessary Mathematica statement and its output are

\[
\text{In}[3] := \text{soln1} = x^2 \text{Simplify}[\ %/x^\text{\[Gamma]} ] \\
\text{Out}[3] = (-1 + \gamma) a_0 + x \gamma(1 + \gamma) a_1 + x^2(a_0 + (2 + 3 \gamma + \gamma^2)a_2 + \ldots)
\]

We have written out only the terms with the lowest powers of \(x\) for the moment because they are the terms on which we must focus first. Since this series must be identically zero for all \(x\) and since different powers of \(x\) are linearly independent, the series will sum to zero only if the coefficient of each power of \(x\) is separately zero. To simplify the application of this criterion, we might extract the coefficients themselves as a list with the statement

\[
\text{In}[4] := \text{coefs} = \text{CoefficientList}[\ \text{soln1}, x] \\
\text{Out}[4] = \{-1 + \gamma \gamma a_0, \gamma(1 + \gamma) a_1, a_0 + (2 + 3 \gamma + \gamma^2)a_2, \ldots\}
\]

Following the method of Frobenius, we determine acceptable values of \(\gamma\) by requiring that the coefficient of the lowest power of \(x\) (i.e., the first element in this list) be zero under the restriction that \(a_0 \neq 0\). The statements

\[
\text{In}[5] := \text{coefs}[[1]]/a[0] \\
\text{Out}[5] = \{-1 + \gamma \gamma\} \\
\text{In}[6] := \text{gam} = \text{Solve}[\ % == 0, \\text{\[Gamma]} ] \\
\text{Out}[6] = \{\{\gamma \rightarrow 0\}, \{\gamma \rightarrow 1\}\}
\]

reveal that \(\gamma\) can be either 0 or 1. When \(\gamma = 1\), the coefficients of successive powers of \(x\) are

\[
\text{In}[7] := \text{coefs1} = \text{coefs} /. \text{gam}[[2]] \\
\text{Out}[7] = \{0, 2a_1, a_0 + 6a_2, a_1 + 12a_3, a_2 + 20a_4, a_3 + 30a_5, a_4 + 42a_6, a_5 + 56a_7, a_6 + 72a_8, a_7, a_8\}
\]

Each of these coefficients must itself be zero. Ignoring the first coefficient (which is already zero) and the last two coefficients (which are incomplete because of our truncation of the series), we extract a set of eight simultaneous linear equations and solve them for the coefficients with the statements

\[
\text{In}[8] := \text{eqs1} = \text{Table}[\ \text{coefs1}[[i+1]] == 0, \{i, 8\}] \\
\text{Out}[8] = \{2a_1 == 0, a_0 + 6a_2 == 0, a_1 + 12a_3 == 0, \ldots\} \\
\text{In}[9] := \text{vars1} = \text{Table}[\ a[i], \{i, 8\}] \\
\text{Out}[9] = \{a_1, a_2, a_3, \ldots\}
\]

\textit{\textsuperscript{21}To shorten the expression, we use the symbol \(a_i\) rather than \(a[i]\) for the \(i\)-th coefficient in the original series.}
11.5. SOLVING ODES SYMBOLICALLY WITH MATHEMATICA

In[10]:= Solve[ eqs1, vars1 ][[1]]
Out[10]= \{a_1 \to 0, a_2 \to -\frac{a_0}{6}, a_3 \to 0, a_4 \to \frac{a_0}{120}, a_5 \to 0, a_6 \to -\frac{a_0}{5040}, a_7 \to 0, a_8 \to \frac{a_0}{362880}\} 

Finally, we can substitute these solutions into the original series with the statement

In[11]:= y[x] /. \[Gamma] \to 1;
In[12]:= % /. %10;
In[13]:= Collect[ %, a[0] ]
Out[13]= (x - x^3/6 + x^5/120 - x^7/5040 + x^9/362880) a_0

Recognizing that 3! = 6, 5! = 120, 7! = 5040, and 9! = 362880, we then also recognize that the series in parentheses is in fact the beginnings of the series for \(\sin(x)\) and this solution apparently is \(a_0 \sin(x)\).

The solution when \(\gamma = 0\) can be found in much the same way. We begin by invoking the statement

In[14]:= coefs0 = coefs /. gam[[1]]
Out[14]= \{0, 0, a_0 + 2 a_2, a_1 + 6 a_3, a_2 + 12 a_4, a_3 + 20 a_5, a_4 + 30 a_6, a_5 + 42 a_7, a_6 + 56 a_8, a_7, a_8\}

to determine the coefficients for this case. As before, each of these coefficients must itself be zero. Ignoring the first two coefficients (which are already zero) and the last two (which are incomplete because of our truncation of the series), we extract a set of seven simultaneous linear equations and solve them for the coefficients with the statements

In[15]:= eqs0 = Table[ coefs0[[i+2]] == 0, {i, 7} ]
Out[15]= \{a_0 + 2 a_2 == 0, a_1 + 6 a_3 == 0, a_2 + 12 a_4 == 0, a_3 + 20 a_5 == 0, ..., \}
In[16]:= vars0 = Table[ a[i+1], {i, 7} ]
Out[16]= \{a_2, a_3, a_4, ..., \}
In[17]:= Solve[ eqs0, vars0 ][[1]]
Out[17]= \{a_2 \to -\frac{a_0}{2}, a_3 \to -\frac{a_1}{6}, a_4 \to \frac{a_0}{24}, a_5 \to \frac{a_1}{120}, a_6 \to -\frac{a_0}{720}, a_7 \to -\frac{a_1}{5040}, a_8 \to \frac{a_0}{40320}\}

Here, we have noted that the conditions imposed on the solution do not constrain either \(a_0\) or \(a_1\).

Finally, we can substitute these solutions into the original series with the statement

In[18]:= y[x] /. \[Gamma] \to 0;
In[19]:= % /. %17;
In[20]:= Collect[ %, \{a[0], a[1]\} ]
Out[13]= (1 - \frac{x^2}{2} + \frac{x^4}{24} - \frac{x^6}{720} + \frac{x^8}{40320}) a_0 + \left(x - \frac{x^3}{6} + \frac{x^5}{120} - \frac{x^7}{5040}\right) a_1

This value of \(\gamma\) has, in fact, yielded a solution containing two arbitrary constants, exactly what we expect in the general solution to a second-order ODE. Interestingly, the series multiplied by \(a_1\) in this result is identical to the (only) series obtained when \(\gamma\) was set to the value 1, i.e., it is again the function \(\sin(x)\). Recognizing that 2! = 2, 4! = 24, 6! = 720, and 8! = 40320, we recognize that the other series in this result—the one multiplied by \(a_0\)—is in fact the beginnings of the series for \(\cos(x)\). Thus, the solution obtained by the method of Frobenius is apparently \(a_0 \cos(x) + a_1 \sin(x)\), where \(a_0\) and \(a_1\) are to be determined by initial or boundary conditions.
11.6 Algorithms for Solving ODEs Numerically

A numerical solution to an ordinary differential equation emerges from the application of a procedure—frequently called an algorithm—for calculating approximate values of the dependent variables at a succession of values of the independent variable. All numerical methods for solving ordinary differential equations exploit the fact that the differential equations determine the rates of change of the dependent variables from the dependent variables themselves. Because solutions obtained numerically are approximate, we must give attention not only to the methods themselves but also to means by which we can assess the accuracy of the solutions obtained.

For the sake of a simple discussion, we shall, in laying out the essence of each of several algorithms, suppose that we are dealing with a single first-order ODE and an initial condition of the form

$$\frac{dx}{dt} = f(x, t) \quad ; \quad x(0) = x_0$$

(11.76)

where \( f(x, t) \) and \( x_0 \) are known from the beginning. The relatively straightforward extension of the initial discussion to systems of first-order equations and to single equations of higher order will be illustrated in several specific examples but will not be explicitly discussed in general terms.

11.6.1 Euler’s Method

Euler’s method, which embodies the simplest numerical approach to ODEs, is based on the assumption that the rates of change of the dependent variables do not themselves change very quickly. Thus, given a short enough time interval \( \Delta t \), the rates of change throughout that interval may be regarded, at least approximately, as constant and equal to the rates of change at the beginning of the interval. For example, provided \( \Delta t \) is not too large,\(^\text{22}\) we can write the approximation

$$\frac{dx}{dt} = f(x, t) \quad \Rightarrow \quad \frac{x(t + \Delta t) - x(t)}{\Delta t} \approx f(x, t) \quad \Rightarrow \quad x(t + \Delta t) \approx x(t) + f(x, t) \Delta t$$

(11.77)

The value \( x(t + \Delta t) \) of the dependent variable at \( t + \Delta t \) is the value \( x(t) \) at time \( t \) plus the amount \( \Delta x \) by which \( x \) changes in the interval, where we estimate \( \Delta x \) by multiplying the rate of change of \( x \) given by \( f(x(t), t) \) at the beginning of the interval by the elapsed time \( \Delta t \), i.e., \( \Delta x = f(x(t), t) \Delta t \). In an alternative notation, if we think of \( t \) as the “old” time and \( t + \Delta t \) as the “new” time, we might express the basic stepping equations as

$$f_{\text{old}} = f(x_{\text{old}}, t_{\text{old}}) \quad ; \quad x_{\text{new}} = x_{\text{old}} + f_{\text{old}} \Delta t \quad ; \quad t_{\text{new}} = t_{\text{old}} + \Delta t$$

(11.78)

Starting from the initial condition as the first \( x_{\text{old}} \) (and a choice of time step \( \Delta t \)—see later), we can then use these stepping equations repeatedly to move from knowledge of \( x(0) \) to knowledge of \( x(\Delta t) \) to knowledge of \( x(2 \Delta t) \) to . . ., continuing as long as our patience endures. The basic strategy of Euler’s method would be summarized in the algorithm in Table 11.1.\(^\text{23}\) As this example shows, we can generate a complete—though approximate—solution from initial knowledge of only

1. The differential equation, which determines the rate of change of the dependent variable from the dependent and independent variables themselves,

2. The initial condition, which starts the process by providing the first row in a table containing \( t, x, \) and \( f \),

\(^{22}\)The precise meaning of “too large” is difficult to define. In general terms, if \( T \) is a typical time during which the solution changes appreciably (say by 10–20%), then a value of \( \Delta t \) satisfying \( \Delta t < T \) will probably yield an adequate solution. Each case must be examined on its own terms; no general rules can be formulated.

\(^{23}\)In many cases, the ultimate and penultimate pairs of statements in the loop can be combined into a single pair of statements. We here refrain from that more compact expression so that the two distinct operations—calculating the new to complete the step and replacing the old with the new to prepare for the next step—will remain distinct.
3. Specific values for any parameters—here there happen to be none—in the differential equation, and
4. A choice of time step $\Delta t$.

With this information as input, each pass through the loop in the above algorithm generates a new row in the table containing $t$, $x$, and $f$.

To be even more concrete (and to illustrate the simple extension to a system of ODEs), let us work out by hand the first few steps in the Euler solution of Eq. (11.19) describing a chain radioactive decay. Reflecting the differential equations, the stepping equations for this specific case are

\begin{align*}
A(t + \Delta t) &= A(t) + \frac{dA}{dt}(t) \Delta t = A(t) - k_A A(t) \Delta t \quad \text{(11.79)} \\
B(t + \Delta t) &= B(t) + \left[k_A A(t) - k_B B(t)\right] \Delta t \quad \text{(11.80)} \\
C(t + \Delta t) &= C(t) + k_B B(t) \Delta t \quad \text{(11.81)}
\end{align*}

For definiteness,\textsuperscript{24} we take $A(0) = 1000.000$, $B(0) = C(0) = 0.000$, and $k_A = k_B = 0.100$, and we select a time step of $\Delta t = 0.250$.\textsuperscript{25} Equations (11.79), (11.80), and (11.81) at $t = 0.0$ then become

\begin{align*}
A(0.00 + 0.25) &= 1000.0 - 0.1 \cdot 1000.0 \cdot 0.25 = 975.000 = A(0.25) \\
B(0.00 + 0.25) &= 0.0 + \left[0.1 \cdot 1000.0 - 0.1 \cdot 0.0\right] \cdot 0.25 = 25.000 = B(0.25) \\
C(0.00 + 0.25) &= 0.0 + 0.1 \cdot 0.0 \cdot 0.25 = 0.000 = C(0.25)
\end{align*}

Continuing the algorithm further by applying Eqs. (11.79), (11.80), and (11.81) for $t = 0.25$, we find that

\begin{align*}
A(0.25 + 0.25) &= 975.0 - 0.1 \cdot 975.0 \cdot 0.25 = 950.625 = A(0.50) \\
B(0.25 + 0.25) &= 25.0 + \left[0.1 \cdot 975.0 - 0.1 \cdot 25.0\right] \cdot 0.25 = 48.750 = B(0.50) \\
C(0.25 + 0.25) &= 0.0 + 0.1 \cdot 25.0 \cdot 0.25 = 0.625 = C(0.50)
\end{align*}

\textsuperscript{24}One disadvantage of numerical approaches is that they are, indeed, numerical. We cannot find solutions containing symbols representing parameters. We must seek solutions for specific numerical values. If we need to know the dependence of the solution on a particular parameter, we will have to generate a separate solution for each desired value of the parameter.

\textsuperscript{25}This choice at the moment is more for convenience than accuracy. The accuracy of the resulting solution is, of course, markedly influenced by the choice of $\Delta t$. We shall return to assess the suitability of this choice in later sections.
Table 11.2: Evolution of three-species radioactive decay determined by Euler’s method with $\Delta t = 0.25$.

<table>
<thead>
<tr>
<th>$t$</th>
<th>$A$</th>
<th>$B$</th>
<th>$C$</th>
<th>$A + B + C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>1000.00</td>
<td>0.000</td>
<td>0.000</td>
<td>1000.00</td>
</tr>
<tr>
<td>0.25</td>
<td>975.000</td>
<td>25.000</td>
<td>0.000</td>
<td>1000.00</td>
</tr>
<tr>
<td>0.50</td>
<td>950.625</td>
<td>48.750</td>
<td>0.625</td>
<td>1000.00</td>
</tr>
<tr>
<td>0.75</td>
<td>926.859</td>
<td>71.297</td>
<td>1.844</td>
<td>1000.00</td>
</tr>
</tbody>
</table>

Going yet one more step, Eqs. (11.79), (11.80), and (11.81) for $t = 0.50$ yield that

$$A(0.50 + 0.25) = 950.625 - 0.1 \cdot 950.625 \cdot 0.25 = 926.859 = A(0.75)$$
$$B(0.50 + 0.25) = 48.750 + \left[0.1 \cdot 950.625 - 0.1 \cdot 48.750\right] \cdot 0.25 = 71.267 = B(0.75)$$
$$C(0.50 + 0.25) = 0.625 + 0.1 \cdot 48.750 \cdot 0.25 = 1.844 = C(0.75)$$

The resulting values are compiled in Table 11.2, in which the first row is provided by the initial conditions and all subsequent rows are determined by the algorithm outlined in Eqs. (11.79), (11.80), and (11.81). Knowing that an exact solution will reflect the conservation of $A + B + C$ [see Eq. (11.20)], we have added a column containing the sum of the amounts of all three species present. Clearly, carried this far, anyway, the solution appears automatically to satisfy that requirement.\(^{26}\)

### 11.6.2 Improved Euler Method

While Euler’s method is simple to motivate, describe, and implement, it unfortunately yields only a coarse approximation. Typically it will require a very small time step (which translates into a large amount of computational time and the potential accumulation of computer roundoff error) to achieve adequate accuracy. Considerable effort has been spent in devising alternative, more refined algorithms that converge more rapidly with a minimal amount of computational labor (and hence less internal roundoff error). The improved Euler method, for example, takes the Euler result at each step to be only an estimate (prediction) of the solution for that step and uses that estimate to refine the solution before going on to the next step. Starting with the values $x_{\text{old}}$ and $t_{\text{old}}$ and a chosen time step $\Delta t$, we

1. calculate $f_{\text{old}}$,

2. calculate the predicted values

$$x_{\text{pred}} = x_{\text{old}} + f_{\text{old}} \Delta t \quad ; \quad t_{\text{new}} = t_{\text{old}} + \Delta t$$

(11.85)

3. calculate $f_{\text{pred}} = f(x_{\text{pred}}, t_{\text{new}})$, and

4. calculate the final (corrected) value

$$x_{\text{new}} = x_{\text{old}} + \frac{1}{2} \left( f_{\text{old}} + f_{\text{pred}} \right) \Delta t$$

(11.86)

\(^{26}\)Actually, however, that the sum $A + B + C$ preserves its initial value is not really a check on the accuracy of the method in this example. If we simply add the three stepping equations we are using [Eqs. (11.79)–(11.81)], we find that $A(t + \Delta t) + B(t + \Delta t) + C(t + \Delta t) = A(t) + B(t) + C(t)$. Solutions generated by Euler’s method applied to this problem automatically satisfy the conservation law, regardless of $\Delta t$. 

11.6. ALGORITHMS FOR SOLVING ODES NUMERICALLY

Table 11.3: Improved Euler algorithm.

\[
\begin{align*}
&dt \leftarrow \Delta t \\
&t_{old} \leftarrow 0.0 \\
x_{old} \leftarrow x_0 \\
\text{loop} \\
&\quad \text{fold} \leftarrow f(x_{old}, t_{old}) \\
&\quad \text{print, t}_{old}, x_{old}, \text{fold} \\
&\quad \text{exit when done} \\
&\quad x_{pred} \leftarrow x_{old} + \text{fold} \cdot dt \\
&\quad t_{new} \leftarrow t_{old} + dt \\
&\quad f_{pred} \leftarrow f(x_{pred}, t_{new}) \\
&\quad x_{new} \leftarrow x_{old} + \frac{1}{2} (\text{fold} + f_{pred}) \cdot dt \\
&\quad x_{old} \leftarrow x_{new} \\
&\quad t_{old} \leftarrow t_{new} \\
&\text{end_loop}
\end{align*}
\]

Choose time step.
Initialize \( t_{old} \).
Initialize \( x_{old} \).
Evaluate \( f_{fold} \).
Display results.
Calculate predicted value.
Evaluate \( f_{pred} \).
Calculate final (corrected) value.
Replace old values with new.

Table 11.4: Evolution of three-species radioactive decay determined by the improved Euler method with \( \Delta t = 0.25 \).

<table>
<thead>
<tr>
<th>( t )</th>
<th>( A )</th>
<th>( B )</th>
<th>( C )</th>
<th>( A + B + C )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>1000.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1000.000</td>
</tr>
<tr>
<td>0.25</td>
<td>975.313</td>
<td>24.375</td>
<td>0.312</td>
<td>1000.000</td>
</tr>
<tr>
<td>0.50</td>
<td>951.234</td>
<td>47.547</td>
<td>1.219</td>
<td>1000.000</td>
</tr>
<tr>
<td>0.75</td>
<td>927.751</td>
<td>69.559</td>
<td>2.690</td>
<td>1000.000</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

In this last step, we in effect estimate the average rate of change over the interval from \( t \) to \( t + \Delta t \) as the average of (a) its value \( f_{old} \) at the beginning of the interval and (b) the best estimate \( f_{pred} \) we have of its value at the end of the interval. In so doing, we admit that the rate of change may change in the interval. Intuitively, for a given \( \Delta t \), the improved Euler method will be more accurate than Euler’s method, which presumes that the average rate of change over the interval is adequately approximated by its value at the beginning of the interval.\(^{27}\)

A full laying out of this improved algorithm differs from the algorithm presented in the previous section in only a few lines, having the expression laid out in Table 11.3. As with Euler’s method, knowledge of the differential equation, the initial condition, and any parameters, together with a choice of a time step, starts a process that leads to (approximate) knowledge of the solution indefinitely into the future.

By way of example (but leaving the arithmetic to the reader), we present in Table 11.4 the results of applying the improved Euler method for the example treated in the previous section. Values in this table should be compared with those in Table 11.2.

\(^{27}\)Phrased as we have described it, the improved Euler method entails first calculating a predicted value at the end point of an interval and then uses that value to determine a corrected final value at that point before going on to the next point. Such methods are often called predictor-corrector methods.
11.6.3 Runge-Kutta Methods

A popular alternative viewpoint, which leads to the deduction of what are called Runge-Kutta stepping algorithms, requires at base that two different Taylor series expansions of the solution match to a chosen number of terms. On the one hand, we know that

\[ x(t + \Delta t) = x(t) + \frac{dx(t)}{dt} \Delta t + \frac{1}{2} \left( \frac{d^2x(t)}{dt^2} \right) \Delta t^2 + O(\Delta t^3) \]

Motivated by the improved Euler method on the other hand, we are tempted to suppose that we might build a stepping algorithm by\(^{28}\)

1. introducing a judicious time \( t' = t + a \Delta t \), where \( a \), constrained by \( 0 \leq a \leq 1 \), is yet to be specified,

2. introducing a tentative solution \( x' = x(t) + b \Delta t f(x(t), t) \), where \( b \), constrained by \( 0 \leq b \leq 1 \), is yet to be specified, and

3. taking \( x(t + \Delta t) = x(t) + \left( w_1 f(x(t), t) + w_2 f(x', t') \right) \Delta t \), where the weights \( w_1 \) and \( w_2 \) are yet to be specified.

The essential idea of the Runge-Kutta approach is to expand the assumption of item 3 as a power series in \( \Delta t \) and require that its first few terms agree with those in the series expressed in Eq. (11.87). To deduce this second Taylor series, we first invoke the two-dimensional Taylor series to find that

\[ f(x', t') = f \left( x + b \Delta t f(x, t), t + a \Delta t \right) \]

\[ = f(x, t) + \frac{\partial f(x, t)}{\partial x} b \Delta t f(x, t) + \frac{\partial f(x, t)}{\partial t} a \Delta t + O(\Delta t^2) \]  \hspace{1cm} (11.88)

where—noting at item 3 above that this term will ultimately be multiplied by \( \Delta t \)—we have included only terms through those first order in \( \Delta t \). Then, we find that

\[ x(t + \Delta t) = x(t) + w_1 f(x(t), t) \Delta t \]

\[ + w_2 \left( f(x, t) + \frac{\partial f(x, t)}{\partial x} b \Delta t f(x, t) + \frac{\partial f(x, t)}{\partial t} a \Delta t + O(\Delta t^2) \right) \Delta t \]

\[ = x(t) + (w_1 + w_2) f(x(t), t) \Delta t \]

\[ + \left( w_2 b f(x, t) \frac{\partial f(x, t)}{\partial x} + w_2 a \frac{\partial f(x, t)}{\partial t} \right) \Delta t^2 + O(\Delta t^3) \]  \hspace{1cm} (11.89)

which will match Eq. (11.87) through terms of order \( O(\Delta t^2) \) if we choose

\[ w_1 + w_2 = 1 \quad a w_2 = b w_2 = \frac{1}{2} \quad \text{or} \quad w_1 + w_2 = 1 \quad a = b \quad w_2 = \frac{1}{2a} \]  \hspace{1cm} (11.90)

\(^{28}\)Note that the steps laid out here reduce to the improved Euler method if we take \( a = b = 1 \) and \( w_1 = w_2 = 1/2 \).
### Table 11.5: A Runge-Kutta Method.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta t$</td>
<td>$\Delta t$</td>
</tr>
<tr>
<td>$t_{old}$</td>
<td>0.0</td>
</tr>
<tr>
<td>$x_{old}$</td>
<td>$x_0$</td>
</tr>
<tr>
<td>$f_{old}$</td>
<td>$f(x_{old}, t_{old})$</td>
</tr>
<tr>
<td>$t_{new}$</td>
<td>$t_{old} + \Delta t$</td>
</tr>
<tr>
<td>$x_{new}$</td>
<td>$x_{old} + \frac{(k1 + k2)}{2}$</td>
</tr>
<tr>
<td>$t_{old}$</td>
<td>$t_{new}$</td>
</tr>
<tr>
<td>$k1$</td>
<td>$f(x_{old}, t_{old}) \cdot \Delta t$</td>
</tr>
<tr>
<td>$k2$</td>
<td>$f(x_{old} + \frac{k1}{2}, t_{old} + \frac{\Delta t}{2}) \cdot \Delta t$</td>
</tr>
<tr>
<td>$k3$</td>
<td>$f(x_{old} + \frac{k2}{2}, t_{old} + \frac{\Delta t}{2}) \cdot \Delta t$</td>
</tr>
<tr>
<td>$k4$</td>
<td>$f(x_{old} + k3, t_{old} + \Delta t) \cdot \Delta t$</td>
</tr>
<tr>
<td>$x_{new}$</td>
<td>$x_{old} + \frac{(k1 + 2k2 + 2k3 + k4)}{6}$</td>
</tr>
</tbody>
</table>

Choose time step.  
Initialize $t_{old}$.  
Initialize $x_{old}$.  
Evaluate $f_{old}$.  
Display results.  
Find change using rate at $t_{old}$.  
Find change using rate at $t_{new}$.  
Calculate $t_{new}$.  
Calculate $x_{new}$.  
Replace old values with new.

We have discovered a multitude of Runge-Kutta schemes, one corresponding to each possible choice of $a$, $b$, $w_1$, and $w_2$. One common choice embodies the values in footnote 28—values that reveal that the improved Euler method is a member of this Runge-Kutta family. A second common choice takes $w_1 = 0$, $w_2 = 1$, and $a = b = \frac{1}{2}$, in which case the stepping equations become

$$ t' = t + \frac{1}{2} \Delta t \quad ; \quad x' = x + \frac{1}{2} f(x, t) \Delta t \quad ; \quad x(t + \Delta t) = x(t) + f(x', t') \Delta t $$ (11.91)

With this scheme, we first step half way over the interval with Euler’s method, then we use that result to estimate the rate of change at the midpoint of the interval and, finally, we use the rate of change at the midpoint to project the solution at the end of the interval. Clearly, this route provides yet another means to estimate the average rate of change over the interval; it is called the midpoint method.

Even though the Runge-Kutta algorithm for $w_1 = w_2 = \frac{1}{2}$ and $a = b = 1$ coincides with the improved Euler method, it is usually presented in the form shown in Table 11.5. This form is more compatible with the most convenient expressions of other algorithms in the broad Runge-Kutta family.

The algorithm described in the previous two paragraphs is known as a second-order Runge-Kutta algorithm because its deduction entailed matching terms in two Taylor expansions through those of order $\Delta t^2$. We could, of course, replace the expression in item 3 above with a more elaborate expression, expand it to include higher-order terms in $\Delta t$, and insist on agreement with a higher-order Taylor expansion deduced from the differential equation. The calculational labor becomes increasingly complicated. Partly because of its popularity, we present without derivation the essence of a fourth-order Runge-Kutta algorithm, limiting ourselves only to the steps that would replace the ones calculating $k1$, $k2$, and $x_{new}$ in the above algorithm:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k1$</td>
<td>$f(x_{old}, t_{old}) \cdot \Delta t$</td>
</tr>
<tr>
<td>$k2$</td>
<td>$f(x_{old} + \frac{k1}{2}, t_{old} + \frac{\Delta t}{2}) \cdot \Delta t$</td>
</tr>
<tr>
<td>$k3$</td>
<td>$f(x_{old} + \frac{k2}{2}, t_{old} + \frac{\Delta t}{2}) \cdot \Delta t$</td>
</tr>
<tr>
<td>$k4$</td>
<td>$f(x_{old} + k3, t_{old} + \Delta t) \cdot \Delta t$</td>
</tr>
<tr>
<td>$x_{new}$</td>
<td>$x_{old} + \frac{(k1 + 2k2 + 2k3 + k4)}{6}$</td>
</tr>
</tbody>
</table>

This fourth-order algorithm emerges when the two Taylor series are matched through the terms involving $\Delta t^4$. A higher-order method, of course, is more accurate than a lower-order method for a

---

29 Actually, we choose only $a$. Then $b = a$, $w_2 = 1/2a$, and $w_1 = 1 - w_2$ are fixed.
CHAPTER 11. SOLVING ORDINARY DIFFERENTIAL EQUATIONS

given time step. Equivalently, a lower-order method requires a smaller time step to give the same
degree of accuracy as a higher-order method.

11.6.4 Assessing Accuracy

Numerical evaluations, of course, only approximate the solution to ordinary differential equations. Thus, we cannot complete a solution without also assessing its accuracy. Furthermore, this task
must be accomplished without knowledge of the exact solution. The importance of being aware that
numerical methods are always approximate cannot be overstressed.

Two distinctly different sorts of errors can occur. Truncation errors arise because the solution
has been based on a finite-difference approximation to the derivatives appearing in the equation; roundoff errors arise because computers do not store non-integers to 100% precision and, in an
iterated calculation where each step depends on the previous step, the imprecision with which each
component is represented within the computer can accumulate as the number of steps increases. Truncation errors become smaller as the step size is reduced. Roundoff errors, unfortunately, become
more significant as step size is reduced (because, with smaller steps, more arithmetic must be done).
Usually, roundoff errors are negligible, the more so as the sophistication of the algorithm increases
(and, hence, the amount of arithmetic decreases). Provided we do not strive for accuracy greater
than about 1 part in $10^5$ (with single precision floating point arithmetic), we can usually ignore
roundoff errors. Thus, provided the solution we seek does not vary too rapidly on the time scale
defined by the time step in use, the quickest way to obtain a reasonably reliable estimate of
truncation error is to solve the equation with two different step sizes, the second being half of the
first, and compare the two results. Presuming that roundoff error has not begun to be important,
we can be confident that the second result is more accurate than the first. Thus, if the two agree
to 1 part in $10^3$, say, we can with reasonable confidence assume that the second value is good to one
part in $10^3$. Indeed, the second value is probably better than that, but assessing its accuracy by this
method would entail obtaining a third value by using a step size half of that used to determine the
second value. Indeed, one strategy for achieving a desired accuracy with reasonable certainty is to
solve the ODE repeatedly by a particular method, halving the step size each time, and continuing
until the new value received differs from its predecessor by less than the desired accuracy (though
we must be careful not to push this approach so far that roundoff problems within the computer
begin to become significant).

We will illustrate this approach in the context of a specific example as soon as we are ready to use the computer to do the arithmetic.

From a more sophisticated perspective, numerical analysts have deduced expressions for the
error in various approaches to solving ODEs numerically. To assess the error in Euler’s method, for
example, we begin by noting the Taylor theorem with remainder, which asserts that

$$x(t + \Delta t) = x(t) + \frac{dx}{dt}(t) \Delta t + \frac{1}{2} \frac{d^2 x}{dt^2}(\xi) \Delta t^2$$

(11.92)

where $\xi$ is a value in the interval $t \leq \xi \leq t + \Delta t$. This expression is exact, though it is only somewhat useful because it tells us only the order of magnitude of the error; it gives no clue as to the actual
value of $\xi$. Nonetheless, we can conclude that

$$x_{\text{exact}}(t + \Delta t) = x_{\text{Euler}}(t + \Delta t) + \frac{1}{2} \frac{d^2 x}{dt^2}(\xi) \Delta t^2$$

(11.93)

or that

$$|x_{\text{exact}}(t + \Delta t) - x_{\text{Euler}}(t + \Delta t)| \leq \text{Max} \left( \left| \frac{1}{2} \frac{d^2 x}{dt^2}(\xi) \right| \right) \Delta t^2$$

(11.94)

which provides yet another motivation for casting problems in dimensionless form.

This assertion is refined in the next paragraphs.

Warning: This approach is not entirely secure if the procedure converges slowly as the time step is reduced. Remember that $\sum_{n=1}^{\infty} 1/n$ diverges even though the effect of the millionth term added to the partial sum starts in the sixth decimal place.
Thus, we learn that, with Euler’s method, the (truncation) error per step varies as the square of the step size; halving the step size will reduce the error per step by a factor of four.

Statements similar to Eq. (11.93) can be deduced for all of the methods that we have described in this section. Without proof, we present the properties

\[
x_{\text{exact}}(t + \Delta t) = x_{\text{Euler}}(t + \Delta t) + O(\Delta t^2) \quad (11.95)
\]

\[
x_{\text{ImpEul}}(t + \Delta t) + O(\Delta t^3) \quad (11.96)
\]

\[
x_{\text{RK2}}(t + \Delta t) + O(\Delta t^3) \quad (11.97)
\]

\[
x_{\text{RK4}}(t + \Delta t) + O(\Delta t^5) \quad (11.98)
\]

For these four common methods, halving the step size reduces the error per step by a factor of four, eight, eight, and thirty-two, respectively. Further, on the basis of these relationships, we characterize “Euler” as a first-order method, “ImpEul” and “RK2” as second-order methods, and “RK4” as a fourth-order method because their derivations involve matching Taylor series to include terms in $\Delta t$, $\Delta t^2$, $\Delta t^3$, and $\Delta t^4$, respectively.

Unfortunately, the analysis described briefly in the previous paragraph is not the whole story. Each of the statements in that paragraph is correct, provided we assume that the step that arrives at the various estimates of $x(t + \Delta t)$ starts with the exact solution at time $t$. Except for the first step, which moves away from (exact) initial conditions, that assumption is invalid. Truncation errors per step compound as more and more steps are taken, and a full assessment of the error in a solution obtained by one or another numerical means must recognize this cumulation. Beyond the error per step, we must be aware of the global truncation error, which attempts to estimate how much error accumulates in the course of working out a solution over the entire desired range of the independent variable. The task of assessing global error is extremely difficult. Crudely, however, if we simply add up expressions like those in the previous paragraph for the $N$ steps in an entire solution and suppose—without much justification—that each step contributes about the same amount, we would conclude that the global truncation error would be order $N O(\Delta t^p)$ when the error per step is of order $O(\Delta t^p)$. For a fixed interval, $N$ is itself of order $1/\Delta t$. Thus, we infer that the global truncation error is of order $O(\Delta t^{p-1})$. This result does not help us much in determining the global error. It does, however, support the conclusion that, for the four methods in the previous paragraph, halving the step size will reduce the global error by a factor of two, four, four, and sixteen, respectively. In particular, halving the step size with the fourth-order Runge-Kutta method will add at least one more decimal digit to the accuracy of the solution overall (and may well do much better).

We have already mentioned that conserved quantities like energy, linear momentum, and angular momentum can sometimes also be used as a check on the accuracy of an evolving solution. Whenever such a conserved quantity exists, its initial value must, of course, be preserved (within some limits). Failure of a particular solution to conform to that requirement signals a need for a smaller time step or a more sophisticated algorithm. Note, however, that some algorithms (e.g., the Euler algorithm for the three-species radioactive decay) automatically preserve one or more conserved quantities; preservation in such cases does not provide any information about the accuracy of the solution itself.

### 11.6.5 Adaptive Methods

In the previous subsections, we assumed that the user of a particular algorithm would actually view the values obtained for different time steps and decide personally when to stop by examining the changes that occur as the time step is successively halved. One can, of course, program a computer to make those decisions. In essence, the program generates the solution with one time step, then

---

repeatedly generates it with a succession of ever smaller time steps, comparing each new solution
with its immediate predecessor and stopping when the absolute value of the difference is smaller than
a tolerance—either absolute or relative—prescribed in advance. As a guard against an infinite loop,
these algorithms should also stop if the desired tolerance has not been achieved in some maximum
number of refinements and should print a warning when the desired tolerance has not in its judgment
been achieved.

Another family of algorithms aims to minimize computational labor by estimating—though the
methods for doing so are often themselves approximate—the accuracy obtained at each step along
the way to a solution. A trial row in the table of data is generated, and the error is assessed. If the
error is within a user-specified tolerance, the program moves on to the next row in the data table. If
the error exceeds the desired tolerance, the program repeats the calculation with progressively smaller
time steps until the desired tolerance is achieved and only then is a new row added to the table.
The procedure also contains means by which the time step is increased when the estimated error is
less than the specified tolerance. Thus, the time step fluctuates as the solution unfolds, being small
when the solution is changing rapidly and large when the solution is changing slowly. Because of
this feature, the procedure is said to be adaptive. Without the overlay of an elaborate interpolation,
adaptive methods have the disadvantage of generating solutions at irregularly spaced times. That
disadvantage, however, is frequently outweighed by the substantial advantage of concentrating the
computational effort in regions where the solution changes rapidly.

11.6.6 Multistep Methods

All of the methods so far discussed have generated the solution at a particular point from knowledge
of the solution at a single earlier point, though many have interpolated solutions at several points
between the initially known point and the desired end point; such methods are called single-step
methods. Some of the solvers available in some software packages use multistep methods, which
reduce the need for the interpolative procedures by projecting the solution at the next point from
knowledge of the solution at several previous points. In seeking a solution to Eq. (11.76), for example,
we would choose a time step $\Delta t$ and introduce the points $t_i = i \Delta t$ and the notation $x_i = x(t_i)$ and
$f_i = f(x_i, t_i)$.

Then, supposing we already had in hand estimates of the solution at four (say) earlier points, we might estimate the solution at $t_{i+1}$ by the formula

$$x_{i+1} = x_i + \frac{\Delta t}{24} \left( 55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3} \right) \quad (11.99)$$

(which represents a particular fourth-order member of a family of methods known as Adams-
Bashforth methods). In effect, we are estimating the average derivative over the interval
t_{i+1} \leq t \leq t_i as a weighted average of the derivatives we can compute at the four points previ-
ous to the one at $t_{i+1}$, all of which we would—of course—have to have in hand at the time we took
this step.

Because they do not require interpolation within the interval over which we are stepping at any
moment, multistep methods can be computationally extremely efficient. Since we cannot use such a
method until we have in hand the solution at several—in the above example four—points, however,
these methods are not self starting. We must adopt some other method to obtain from the initial
values however many values are needed to support the multistep method. To start a solution based
on the stepping formula of Eq. (11.99), we might use a fourth-order Runge-Kutta method to obtain
$x_1$, $x_2$, and $x_3$ from $x_0$. Then we could shift to the fourth-order Adams-Bashforth formula for all
subsequent points.

34 Sometimes the assessment of accuracy will be based not on the generated solution itself but on some property,
e.g., the energy, computed from that solution.

35 In contrast, Euler’s method, for example, would require using throughout a time step small enough to generate
an accurate solution in regions where the solution varies most rapidly. Consequently, too much computational effort
would be focused in regions where the solution changes slowly enough that a larger time step would be acceptable.

We could, of course, simply take the result given by Eq. (11.99) as the solution at \( t_{i+1} \). For a more sophisticated solution, we could regard that result as a prediction and use it in a paired formula to “correct” the prediction. An appropriate fourth-order corrector formula is a member of the Adams-Moulton family, namely

\[
x_{i+1} = x_i + \frac{\Delta t}{24} \left( 9f_{i+1} + 19f_i - 5f_{i-1} + f_{i-2} \right)
\]  

(11.100)

Once a sufficient number of values is in hand, we could at each step invoke Eq. (11.99) to predict the next solution and then Eq. (11.100) to refine it. We might call the resulting method the Adams-Bashforth-Moulton method.

11.9 Solving ODEs Numerically with OCTAVE

Note: All OCTAVE program (.m) files referred to in this chapter are available in the directory $HEAD/octave, where (as defined in the Local Guide) $HEAD must be replaced by the appropriate path for your site. At some sites, this directory or some other directory containing these files may also have been placed in OCTAVE’s default search path. If so, the files can be found by OCTAVE without explicit specification of a path. Otherwise, you will have to use the full path to copy them into your default directory to access them.

ODEs can be solved numerically either by using OCTAVE’s elementary commands or, more simply, by invoking OCTAVE’s built-in routine lsode.

11.9.1 Using Elementary Commands

To carry the process worked out by hand in Section 11.6.1 much beyond the first few steps is a job for a computer. Suppose, for example, we wanted to determine the Euler approximation to the solution of Eq. (11.19) for chain radioactive decay over the time interval \( 0.0 \leq t \leq 50.0 \) with a time step of \( \Delta t = 0.25 \) and \( k_A = k_B = 0.1 \). We could exploit the capabilities of OCTAVE as follows. We anticipate that values will be stored in arrays \( t, A, B, \) and \( C \). Ultimately these arrays will have 201 elements (time interval divided by time step = 50.0/0.25 = 200 steps, but we need an element also for the initial value). Initially, however, we set only the first element of each array and, of course, the parameters and the time step. Possible statements are

```octave
>> t(1) = 0.0; A(1) = 1000.0; % Set initial values.
>> B(1) = 0.0; C(1) = 0.0;
>> kA = 0.1; kB = 0.1; % Set parameters.
>> dt = 0.25; % Set time step.
```

The solution is then calculated, time instant by time instant, with the loop

```octave
>> for i = 2:201
     A(i) = A(i-1) - kA*A(i-1)*dt;
     B(i) = B(i-1) + kA*A(i-1)*dt - kB*B(i-1)*dt;
     C(i) = C(i-1) + kB*B(i-1)*dt;
     t(i) = t(i-1) + dt;
 end
```

which calculates each new \( A, B, \) and \( C \) using the stepping equations given in Eqs. (11.79), (11.80), and (11.81) and each new \( t \) simply by adding the time step \( dt \) to the previous \( t \).\(^{37}\) The simple statement

\(^{37}\)For simplicity, we have calculated each new time by adding the time step to the previous time. If we were more concerned about minimizing roundoff error, we would be better off replacing the last statement in the loop with the statement \( t(i) = (i-1) \cdot dt \), though doing so would increase the execution time. (Multiplication takes more time than addition.)
Figure 11.10: Solution of chain radioactive decay via Euler’s method with $\Delta t = 0.25$. The solid, dotted, and dashed lines show $A(t)$, $B(t)$, and $C(t)$, respectively.

```matlab
>> for i = 1:4
    sprintf('%6.2f %9.3f %9.4f %9.4f %9.2f
', ...
        t(i), A(i), B(i), C(i), A(i)+B(i)+C(i) )
end
ans = 0.00 1000.000 0.0000 0.0000 1000.00
ans = 0.25 975.000 25.0000 0.0000 1000.00
ans = 0.50 950.625 48.7500 0.6250 1000.00
ans = 0.75 926.859 71.2969 1.8438 1000.00
```
displays the resulting solution and the conserved quantity $A + B + C$ for the first few time steps. Reassuringly, the values here agree with those in Table 11.2. The more interpretable graphical output of Fig. 11.10 is produced by the statements

```matlab
>> plot( t, A, 'linewidth', 3.0, 'color', 'black' )
>> title( 'Three-Species Decay Chain', 'fontsize', 16 )
>> xlabel( 'Time', 'fontsize', 14 )
>> ylabel( 'Number of Atoms Present', 'fontsize', 14 )
>> set( gca, 'fontsize', 14 )
>> hold on
>> plot( t, B, 'linewidth', 3.0, 'color', 'black', 'linestyle', ':' )
>> plot( t, C, 'linewidth', 3.0, 'color', 'black', 'linestyle', '--' )
>> hold off
```

To assess the accuracy achieved by Euler’s method more completely, we repeat the above calculation for several different time steps, finding the representative values shown in Table 11.6. The convergence of $A(6.0)$, $B(6.0)$, and $C(6.0)$ as $\Delta t$ is reduced by successive factors of two is apparent.

---

38We have removed several extraneous lines from the output produced by this statement.
Table 11.6: Solution to radioactive decay at $t = 6.0$ for indicated time steps. These solutions were obtained via Euler’s method.

<table>
<thead>
<tr>
<th>$t$</th>
<th>$\Delta t$</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.000</td>
<td>2.0000</td>
<td>512.00</td>
<td>384.00</td>
<td>104.00</td>
</tr>
<tr>
<td>1.0000</td>
<td>531.44</td>
<td>354.29</td>
<td>114.26</td>
<td></td>
</tr>
<tr>
<td>0.5000</td>
<td>540.36</td>
<td>341.28</td>
<td>118.36</td>
<td></td>
</tr>
<tr>
<td>0.2500</td>
<td>544.64</td>
<td>335.16</td>
<td>120.19</td>
<td></td>
</tr>
<tr>
<td>0.1250</td>
<td>546.74</td>
<td>332.20</td>
<td>121.06</td>
<td></td>
</tr>
<tr>
<td>0.0625</td>
<td>547.78</td>
<td>330.73</td>
<td>121.49</td>
<td></td>
</tr>
</tbody>
</table>

Figure 11.11: Solution of chain radioactive decay via Euler’s method. The solid, dotted, and dashed lines show the solution for $\Delta t = 0.0625$, 0.5, and 2.0, respectively. While reducing $\Delta t$ from 2.0 to 0.5 clearly makes a difference on the scale of the graph, the further reduction from 0.5 to 0.0625 is hardly noticeable over the entire range of the independent variable.

Since, even at the end, the approximate solutions still seem to be changing in the units digit with each refinement, we would be off base to claim an accuracy much smaller than $\pm 1.0$ in these values. For some purposes, of course, that accuracy may well be adequate. In particular, to the resolution of the graph in Fig. 11.10, for example, a variation of $\pm 1.0$ in a particular vertical coordinate would hardly show. At the same time, we must remember that these differences may cumulate and may well be larger at $t = 50.0$, for example, than they are at $t = 6.0$. To assess that possibility, we generate the entire solution via Euler’s method for three time steps ($\Delta t = 2.0$, 0.5, and 0.0625), plotting those solutions in Fig. 11.11. The comparison of the resulting graphs is described in the caption to the figure.

A similar demonstration using the improved Euler method is left to an exercise. In particular, the values presented in Table 11.7 reflect the result of that exercise when the procedures leading to Table 11.6 are repeated with the improved Euler method. Clearly, the results in Table 11.7 are converging more quickly on stable values to more decimal places than we observed in Table 11.6.
Table 11.7: Solution to radioactive decay at $t = 6.0$ for indicated time steps. These solutions were obtained via the improved Euler method.

<table>
<thead>
<tr>
<th>$t$</th>
<th>$\Delta t$</th>
<th>$A$</th>
<th>$B$</th>
<th>$C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.0000</td>
<td>2.0000</td>
<td>551.37</td>
<td>322.75</td>
<td>125.88</td>
</tr>
<tr>
<td>1.0000</td>
<td>549.40</td>
<td>327.82</td>
<td>122.78</td>
<td></td>
</tr>
<tr>
<td>0.5000</td>
<td>548.95</td>
<td>328.94</td>
<td>122.10</td>
<td></td>
</tr>
<tr>
<td>0.2500</td>
<td>548.87</td>
<td>329.20</td>
<td>121.95</td>
<td></td>
</tr>
<tr>
<td>0.1250</td>
<td>548.82</td>
<td>329.27</td>
<td>121.91</td>
<td></td>
</tr>
<tr>
<td>0.0625</td>
<td>548.81</td>
<td>329.28</td>
<td>121.90</td>
<td></td>
</tr>
</tbody>
</table>

Table 11.8: A generic function to define ODEs for OCTAVE.

```octave
function Derivs = FunctionName( DepVars, IndVar )
    # Help message (may occupy several lines)
    Intermediate statements.
    ...
    Derivs = [ DerivOfFirstVar; DerivOfSecondVar; ... ]
endfunction
```

Remember that, with Euler’s method, halving $\Delta t$ reduces the global truncation error by a factor of two while, with the improved Euler method, halving $\Delta t$ reduces the global truncation error by a factor of four. Said another way, we would have to reduce $\Delta t$ by a factor of ten to gain one decimal digit in a solution by Euler’s method; we would have to reduce $\Delta t$ by only a factor of $\sqrt{10} = 3.2$ to gain a decimal digit in a solution by the improved Euler method. The values in these two tables reflect this difference.

### 11.9.2 Defining ODEs for OCTAVE

Before using the routine OCTAVE supplies for solving a system of first-order differential equations numerically, we must create an M-file\(^{39}\) whose execution by OCTAVE returns the derivatives of the dependent variables as defined by the differential equations we seek to solve. Fortunately, all routines use the same structure for this M-file. In broad outline, this file will have the form shown in Table 11.8.\(^{40}\) In this structure,

- The first executable line in the function definition must begin with the keyword `function`.
- The function name will be the same as the name of the file in which it is stored. For example, if we choose to call the function for the radioactive decay chain, `decay`, then the file storing the function will be named `decay.m`.
- The variables must be given in the order shown.
- `Derivs`—any available name can be used—is the name of the variable in which the function will return the derivatives of the dependent variables; the returned result must be a vector. We will structure functions to produce column vectors, though row vectors also are accepted.

---

\(^{39}\)See Section 4.9.

\(^{40}\)Remember that comments can be introduced either with `#` or `%`. See item 6 in Section 4.1.
11.9. SOLVING ODES NUMERICALLY WITH OCTAVE

- **DepVars** is a vector that, on input, supplies the values of the dependent variables for which the derivatives are to be computed.
- **IndVar** is a scalar that, on input, supplies the value of the independent variable at which the derivatives are to be computed.
- The final line in the function definition must be the keyword `endfunction`.

Explicit M-files for several examples will be presented in the subsequent subsections.

11.9.3 The Procedure `lsode`

OCTAVE’s procedure for solving ODEs is `lsode`, which is patterned after a large FORTRAN solver (the Livermore Solver for Ordinary Differential Equations), written by Alan Hindmarsh and first appearing in 1980. The simplest statement invoking `lsode` has one of the general forms:

\[
\begin{align*}
\text{soln} &= \text{lsode}( \text{@FunctionName}, \text{InitialConds}, \text{IndepVar} ) \\
[ \text{soln}, \text{istate} ] &= \text{lsode}( \text{@FunctionName}, \text{InitialConds}, \text{IndepVar} ) \\
[ \text{soln}, \text{istate}, \text{msg} ] &= \text{lsode}( \text{@FunctionName}, \text{InitialConds}, \text{IndepVar} )
\end{align*}
\]

where

- **soln**—any legal variable name will do—is the name of the (vector or scalar) variable in which the solution is returned.
- **istate**—any legal variable name will do—is the name of the scalar variable in which a numeric code indicating the success (\(\text{istate} = 2\)) or lack thereof (\(\text{istate} \neq 2\)) of the procedure is returned.
- **msg**—any legal variable name will do—is the name of the variable in which a character string explaining the code in \(\text{istate}\) is returned.
- **FunctionName** is the name of the function defining the equations to be solved; see Section 11.9.2.
- **InitialConds** is the name of the (scalar or vector) variable that provides the initial value of each dependent variable. Alternatively, the scalar or vector value itself may be provided as this argument to `lsode`.
- **IndepVar** is the name of the vector variable that provides the values of the independent variable at which solutions are to be calculated. Alternatively, the vector itself may be provided as this argument to `lsode`.

The procedure is adaptive, which means that, in generating the solution, `lsode` varies the step size dynamically as it discovers how rapidly the solution changes in various regions within the full range of the independent variable. As output, the procedure generates solutions at the user-specified values—usually equally spaced—in \(\text{IndVar}\), but, behind the scenes, \(\text{lsode}\) will be taking steps of various sizes depending on how rapidly the solution changes and the tolerances—default or otherwise—specified when `lsode` is invoked. Because the varying step sizes may not generate solutions at the user-specified values of the independent variable, `lsode` will interpolate when necessary to yield solutions at the requested values of the independent variable.

To speed the actual illustration of the use of `lsode`, we postpone discussing two remaining issues. Sometimes, we will be interested in the dependence of the solution on one or more parameters.

---

\[\text{We shall adhere to providing } \text{FunctionName} \text{ by enclosing the name preceded by an } @ \text{ sign. OCTAVE will also accept the alternative specification that encloses } \text{FunctionName} \text{ in single or double quotation marks. See footnote 36 in Chapter 4.}\]
CHAPTER 11. SOLVING ORDINARY DIFFERENTIAL EQUATIONS

Table 11.9: The OCTAVE function decay.m.

function derivs = decay( n, t )
# DECAY: Returns derivatives for chain radioactive decay.
# The function DECAY defines the rate equations for a three
# species radioactive decay sequence such as
#
# n(1) --kA--> n(2) --kB--> n(3)
#
# kA = 0.1; kB = 0.1;
derivs = [ -kA*n(1); kA*n(1)-kB*n(2); kB*n(2) ];
endfunction

embedded in the differential equations. Unfortunately, lsode does not provide a means to pass
parameters from the program that calls lsode to the function that is called by lsode to define the
differential equations. A means to overcome the absence of this capacity in lsode is laid out at the
end of Section 11.9.4.

Second, the detailed behavior of lsode—method used, tolerances imposed, ...—is controlled
by a number of options, each of which has a (judiciously chosen) default value but each of which
can also be changed should change be warranted in a particular circumstance. For a while, we will
simply accept all default values. Discussion of the value of those defaults and the ways to change
them is postponed to Section 11.9.9.

11.9.4 Chain Radioactive Decay

As a first example of the use of lsode, consider the equations describing chain radioactive decay,
Eq. (11.19). Before we can construct the proper M-file to communicate the differential equations to
OCTAVE’s routine, we must establish a correspondence between the elements of a vector—name it
n—and the dependent variables. Since we have three dependent variables (the number of atoms of
species A, B, and C), n will have three elements, which OCTAVE refers to as n(1), n(2), and n(3).
If we assign the number of atoms of species A to the first element, the number of atoms of species
B to the second element, and the number of atoms of species C to the third element, we have the
associations A mopb n(1), B mopb n(2), and C mopb n(3). Rewriting the differential equations Eq. (11.19)
in terms of the elements of n, we have the system of equations

\[
\frac{d}{dt}n(1) = -k_A n(1) ; \quad \frac{d}{dt}n(2) = k_A n(1) - k_B n(2) ; \quad \frac{d}{dt}n(3) = k_B n(2) \quad (11.101)
\]

Now we are ready to construct the function M-file describing this system of differential equations.
Respecting the general structure described in Section 11.9.2, we would compose the definition shown
in Table 11.9. Here, the first line tells OCTAVE that the function returns the derivatives in the
variable derivs (a vector), that the function is named decay, that the values of the dependent
variables will be supplied in the vector n, and that the value of the the independent variable will
be supplied in the scalar t.\textsuperscript{42} Respectively, the two executable lines define the values of the two
parameters and assign to the value of the function a column vector—semicolons separating entries—
whose elements are the derivatives of the dependent variables as determined by the right hand sides of Eqs. (11.101).

\textsuperscript{42}For this first example, we hard code the values of the parameters kA and kB in the function definition. Later in
this section we will explain how to set these parameters in other ways.
With the function of the previous paragraph stored in the file decay.m in the default directory or in a directory in OCTAVE’s search path, we would produce a solution for the differential equations it defines with the statements

\[
\begin{align*}
&\gg t = [0.0 : 0.25 : 50.0]; & \text{Set times at which solutions are desired.} \\
&\gg ic = [1000.0; 0.0; 0.0 ]; & \text{Set initial conditions.} \\
&\gg [n, istate, msg] = lsode( @decay, ic, t ); & \text{Solve equations at the times in } t.
\end{align*}
\]

The solutions will be generated at the times in \( t \) with the solution at each time placed in the corresponding row of \( n \). Corresponding to the behavior of the FORTRAN program \texttt{LSODE}, the variable \texttt{istate} will be set to a numeric code, which will be 2 if the execution encounters no difficulties, and the variable \texttt{msg} will provide greater detail for the code returned in \texttt{istate}. We have accepted the default tolerances. (See Section 11.9.9.)

Once the solution is in hand in \( n \) and \( t \), we could plot the solution with the simple statement \texttt{plot( t, n )}, which would plot the three columns of \( n \) in a single window, using a different color for each column. For purposes of black and white display on the page, it is better to use the statements

\[
\begin{align*}
&\gg plot( t, n(:,1), 'color', 'black', 'linewidth', 2.0 ) \\
&\gg hold on \\
&\gg plot( t, n(:,2), 'color', 'black', 'linewidth', 2.0, 'linestyle', ':' ) \\
&\gg plot( t, n(:,3), 'color', 'black', 'linewidth', 2.0, 'linestyle', '--' ) \\
&\gg title( 'Three-Species Chain Decay', 'fontsize', 16 ) \\
&\gg xlabel( 'Time', 'fontsize', 14 ) \\
&\gg ylabel( 'Number of Atoms Present', 'fontsize', 14 ) \\
&\gg set( gca, 'fontsize', 14 ) \\
&\gg hold off
\end{align*}
\]

which will distinguish the three variables with line styles rather than color. The resulting graph—Fig. 11.12—shows the time variation of the number of atoms of each species.

Almost always, we will want to explore the dependence of the solution on the values of any parameters, e.g., \( k_A \) and \( k_B \) in the example of this section. Such explorations would be much easier if the parameters could be set before \texttt{lsode} is called and be known inside the function that \texttt{lsode} calls to determine the derivatives. Unfortunately, \texttt{lsode} contains no mechanism to pass parameters to a function that it calls. Instead, we must make use of what are called global variables, which will be defined in the main program and then will be accessible in any subprogram or function that also defines the variables as global. To modify the above example, we would start by editing the function \texttt{decay} to be produce the function defined in Table 11.10. The statement \texttt{global kA kB} in this function instructs the function to look to global storage for the values of \( k_A \) and \( k_B \) to be used. Invocation of \texttt{lsode} by the main program then is modified from the above as follows

\[
\begin{align*}
&\gg global kA kB \\
&\gg kA = 0.1; kB = 0.1; & \text{Declare global variables.} \\
&\gg t = [0.0 : 0.25 : 50.0]; & \text{Set values for global variables.} \\
&\gg ic = [1000.0; 0.0; 0.0 ]; & \text{Set times to be used.} \\
&\gg [n, istate, msg] = lsode( @decaymod, ... \\
&\hspace{1cm} ic, t ); & \text{Set initial conditions.} \\
&\gg plot( t, n ) & \text{Solve equations at the times in } t. \\
& & \text{Plot results.}
\end{align*}
\]

To explore other values for the parameters, we simply reset the value of \( k_A \) and \( k_B \) and re-execute the third, fourth, fifth, and sixth statements in the sequence immediately above. Note, incidentally, that the order and type of the variables stored as global is all that is important; the names of those variables need not be the same in all contexts.

Take a few minutes to explore the effects of different decay constants by trying several different values of \( k_A \) and \( k_B \). Take a few minutes also to compare the graph in Fig. 11.10 with the graph

\footnote{See Section 4.18 for information about the default directory and the search path.}
Figure 11.12: Solution of chain radioactive decay via \texttt{lsode} with a tolerance of about $10^{-8}$ absolute, $10^{-8}$ relative (see Section 11.9.9). The solid, dotted and dashed lines show $A(t)$, $B(t)$, and $C(t)$, respectively.

Table 11.10: The OCTAVE function \texttt{decaymod.m}.

```octave
function derivs = decaymod( n, t )
# DECAYMOD: Returns derivatives for chain radioactive decay.
# The function DECAYMOD defines the rate equations for a three
# species radioactive decay sequence such as
# # n(1) --kA--> n(2) --kB--> n(3)
# # in a way that does not require hard coding of the parameters in
# # the routine. The parameters are passed as global variables.

global kA kB
derivs = [ -kA*n(1); kA*n(1)-kB*n(2); kB*n(2) ];
endfunction
```

in Fig. 11.12. The first of these figures was produced with Euler’s method with a time step of 0.25 and required 201 points in the solution; the second was produced with \texttt{lsode} with the (default) tolerance of $\approx 10^{-8}$ absolute, $\approx 10^{-3}$ relative (see Section 11.9.9). To the resolution of the graphs, the two solutions appear to be pretty much the same. Evidently, the choice of the time step in our application of Euler’s method was adequate to yield a solution accurate to the resolution of the graph, which is the most restrictive statement the evidence we have quoted supports. We might alternatively have inferred the likely appropriateness of the time step $\Delta t = 0.25$ in Euler’s method by noting from the graphs that the solution varies appreciably over a time period on the order of
function derivs = damposc( y, t )
# DAMPOSC: returns derivatives for damped 1D harmonic oscillator
# The function DAMPOSC describes the nondimensionalized equations of
# motion for a damped, one dimensional harmonic oscillator. bb is a
# dimensionless parameter and has a value of bb=b/sqrt(m*k), where b
# is the damping, m is the mass and k is the spring constant.
# bb is passed to damposc as a global variable.

global bb
derivs = [ y(2); -y(1) - bb*y(2) ];
endfunction

seconds. For example, the initial value of $A$ drops by 20% from 1000 to 800 in about 2 or 3 time
units. Since the time step $\Delta t$ is small compared to that characteristic time, we would be justified
in anticipating a reasonably accurate solution.

### 11.9.5 Damped Harmonic Oscillator

Consider next the damped harmonic oscillator described in dimensionless form as a pair of first-order
equations by Eq. (11.30), though we shall for simplicity remove the driving force by setting $\overline{f}(t)$ to
zero. In writing the M-file for this system, we first realize that we have two first-order differential
equations. Thus, the array, which we name $y$, will have two columns and we make the assignments
$x \mapsto y(1)$ and $v \mapsto y(2)$. In these terms, the differential equations become

$$
\frac{d}{dt} y(1) = y(2) ; \quad \frac{d}{dt} y(2) = -y(1) - \beta y(2)
$$

Then, using the symbol $bb$ for $\overline{\beta}$, we create the M-file in Table 11.11 to define the differential
equations. Storing this file in the default directory with the name `damposc.m`, we explore the system
for two different damping constants with the statements

```matlab
>> global bb
>> bb = 0.0;
>> t1 = [0.0 : 0.2 : 30.0];
>> ic = [1.0; 0.0];
>> [x1, istate, msg] = lsode( @damposc, ...
                           ic, t1);
```

```matlab
>> bb = 0.25;
>> t2 = [0.0 : 0.2 : 30.0];
>> ic = [1.0; 0.0];
>> [x2, istate, msg] = lsode( @damposc, ...
                           ic, t2);
```

Finally, we produce graphs showing the position and velocity as functions of time and also the
trajectory in the phase plane ($v$ versus $x$) for each value of the damping constant with the statements

```matlab
>> subplot(2,2,1)
>> plot( t1, x1(:,1), 'linewidth', 2.0, 'color', 'black' )
```
The resulting graphs are shown in Fig. 11.13.

11.9.6 Planetary Orbits

Expressed in dimensionless form in Cartesian coordinates, the equations of motion for a planet of mass \( m \) orbiting a sun of mass \( M \) were presented in Eq. (11.46). We must, however, recast the system of two second-order equations as a quartet of first-order equations. Thus, we view the equations as the system

\[
\begin{align*}
\frac{dx}{dt} &= v_x; & \frac{dv_x}{dt} &= -\frac{x}{(x^2 + y^2)^{\frac{b}{2}}}; & \frac{dy}{dt} &= v_y; & \frac{dv_y}{dt} &= -\frac{y}{(x^2 + y^2)^{\frac{b}{2}}}.
\end{align*}
\] (11.103)

Here, for the sake of generality, we have introduced the parameter \( b \), which will allow us to explore forces other than the inverse square force. For the inverse square force, we simply set the parameter to the value 1.5.

To create the necessary M-file, we set \( x, v_x, y, \) and \( v_y \) into correspondence with the elements \( x(1), x(2), x(3), \) and \( x(4) \) of the four-element vector \( x \). Then, one possibility for the M-file \texttt{planet.m} defining the above equations is shown in Table 11.12. Finally, to calibrate our sense of appropriate initial conditions, we remember from Eq. (11.60) that, in our dimensionless units, a circular orbit is achieved when the speed of the planet is the reciprocal of the square root of the radius of the planet’s orbit. For example, a circular orbit of radius 4.0 units requires a speed of \( \frac{1}{\sqrt{4.0}} = 0.5 \) units.

Now we have all the elements in place to invoke OCTAVE and pursue solutions to the planetary problem. We use \texttt{lsode} again and accept its default tolerances (see Section 11.9.9), we solve the differential equations for an inverse square law force \((b = 1.5)\) twice, first for a circular orbit and then for an elliptical orbit, with the statements

\[
\begin{align*}
\text{>> global b} \\
\text{>> b = 1.5;} \\
\text{>> t = [0.0 : 0.25 : 60.0 ];}
\end{align*}
\]
Figure 11.13: Damped and undamped harmonic oscillators. The upper two graphs show position (solid line) and velocity (dashed line) as functions of time; the lower two graphs show the phase plane plot. Notice the difference between the phase plane orbits.

Table 11.12: The OCTAVE function `planet.m`.

```octave
function derivs = planet(x, t)
    # PLANET: returns the derivatives for planet in field of sun
    # The function PLANET describes the equations of motion for a
    # planet of mass m orbiting a sun of mass M. The parameter b allows
    # the user to explore forces that are not inverse-square. (For an
    # inverse square law, b=1.5.) Entries in the vector of dependent
    # variables are X-position, X-velocity, Y-position, Y-velocity.
    # The parameter must be passed to the routine as a global variable.
    global b
    temp = (x(1)^2 + x(3)^2)^b;
    derivs = [ x(2); -x(1)/temp; x(4); -x(3)/temp ];
endfunction
```
>> ic1=[ 4.0; 0.0; 0.0; 0.5 ];
>> [x1, istate, msg] = lsode( @planet, ic1, t );
>> ic2=[ 4.0; 0.0; 0.0; 0.3 ];
>> [x2, istate, msg] = lsode( @planet, ic2, t );

Figure 11.14 shows the variety of ways the results of this calculation can be displayed. That the first case indeed yields a circular orbit provides some evidence supporting the conclusion that we have generated the solution with adequate accuracy.\footnote{The statements creating the graphs have been omitted but, except for remembering to use the statement \texttt{axis square} to render circles as circles, could be constructed with little difficulty.}

With this example, assessing accuracy by checking the conservation of energy [given by the first member of Eq. (11.53)] and angular momentum [given by the second member of Eq. (11.53)] is worthwhile. The statements

\begin{verbatim}
>> E1 = 0.5*(x1(:,2).^2+x1(:,4).^2) - 1.0./sqrt(x1(:,1).^2+x1(:,3).^2);
>> L1 = x1(:,1).*x1(:,4) - x1(:,3).*x1(:,2);
>> E2 = 0.5*(x2(:,2).^2+x2(:,4).^2) - 1.0./sqrt(x2(:,1).^2+x2(:,3).^2);
>> L2 = x2(:,1).*x2(:,4) - x2(:,3).*x2(:,2);
\end{verbatim}

will calculate these quantities. Then, statements like

\begin{verbatim}
>> E1, L1
\end{verbatim}

will display the resulting values. We find—see your own output—that $E_1 = -0.12500$, $L_1 = 2.0000$, $E_2 = -0.20500$, and $L_2 = 1.2000$ at every point throughout the entire solutions.\footnote{Remember that all calculations in OCTAVE are done in double precision. We can display all digits in the results by executing the statement \texttt{format long}. Of course, the initial values of $E_1$, $L_1$, $E_2$, and $L_2$ remain as already reported, even to double precision. The last values in $E_1$, $L_1$, $E_2$, and $L_2$, respectively, are $-0.12500423982547$, $1.99999660814821$, $-0.205001926089155$, and $1.19998731084075$.}

These results give us no concern at all about the adequacy of the solutions, at least insofar as their respect for known conserved quantities is concerned.

Open orbits, in which a satellite with positive energy moves in the gravitational field of a sun, are also possible. We utilize OCTAVE’s \texttt{for} loop to produce a variety of open orbits for a planet started in different initial positions by executing the statements listed in Table 11.13. Note that, for an orbit that heads directly towards the (attractive) force center ($y_0 = 0$), the numerical solution encounters a divergence, the planet collides with the force center, and (presumably) disappears. That one “orbit” has been omitted from the loop but added with the first statement after the loop. Figure 11.15 shows the resulting output.

11.9.7 Standing Waves in a String

To address a boundary value problem, we must be clever, since the methods available for solving ODEs all suppose that we are dealing with an initial value problem. One strategy (sometimes called a \textit{shooting method}) involves accepting the boundary value at one end of the interval, \textit{guessing} a derivative at that same end, assuming values for any parameters in the equation, solving the resulting \textit{initial value} problem, assessing the extent to which the solution so generated respects the boundary value at the other end of the interval, and repeating the process while tampering with the derivative or, more commonly, with the parameters until satisfactory agreement with both boundary values has been obtained. For standing waves in a string [Eqs. (11.56) and (11.57)] or the quantum particle in an infinitely deep square well [Eq. (11.59)], the expression of the relevant equations as a system of first-order equations is

\begin{equation}
\frac{df}{dx} = g ; \quad \frac{dg}{dx} = -k^2 f ; \quad f(0) = 0 ; \quad f(\ell) = 0 \tag{11.104}
\end{equation}
Figure 11.14: Circular and elliptical orbits for a planet orbiting a sun. The orbits are shown in the $xy$ plane and in two different phase planes.
Table 11.13: OCTAVE statements for open orbits.

\[
\begin{align*}
& \text{>> b=1.5;} \\
& \text{>> axis equal;} \\
& \text{>> set( gca, 'xlim', [-4.0,4.0], 'ylim', [-4.0,4.0], 'fontsize', 14 );} \\
& \text{>> title( 'Open Orbits', 'fontsize', 16 )} \\
& \text{>> xlabel( 'X-Position', 'fontsize', 14 )} \\
& \text{>> ylabel( 'Y-Position', 'fontsize', 14 )} \\
& \text{>> hold on} \\
& \text{>> y0 = [-4.0 -3.0 -2.0 -1.0 1.0 2.0 3.0 4.0];} \\
& \text{>> for i= 1:8} \\
& \quad \text{ic=[-4.0; 1.0; y0(i); 0.0];} \\
& \quad \text{t = [0.0 : 0.25 : 60.0 ];} \\
& \quad [x, istate, msg] = lsode( @planet, ic, t ); \\
& \quad \text{plot( x(:,1), x(:,3), 'linewidth', 2.0, 'color', 'black' );} \\
& \text{end} \\
& \text{>> plot( [-4.0 0.0], [0.0 0.0], 'linewidth', 2.0, 'color', 'black' );} \\
& \text{>> grid on;}
\end{align*}
\]

Table 11.14: The OCTAVE function stdwaves.m.

\[
\begin{align*}
\text{function derivs = stdwaves( f, x )} \\
& \text{# STDWAVES: returns derivatives for standing waves in string} \\
& \text{# The function STDWAVES defines the basic equations for describing} \\
& \text{# standing waves in a string. The parameter must be passed to} \\
& \text{# the routine as a global variable.} \\
& \text{global k} \\
& \text{derivs = [ f(2); -k^2*f(1) ];} \\
& \text{endfunction}
\end{align*}
\]

Thus, setting up the correspondences \( f \mapsto f(1) \) and \( g \mapsto f(2) \), we might define this system for OCTAVE with the M-file shown in Table 11.14 and stored in the default directory with the name stdwaves.m.

Adopting the strategy described in the previous paragraph, we would then accept the requirement that \( f(0) = 0 \), assume a value—say 1.0—for \( df(0)/dx = g(0) \) and a value for \( k \)—say 1.0, generate the solution over \( 0 \leq x \leq 1 \), and see whether it returns to the value \( f(1) = 0 \) (where—because we need numbers to effect a numerical solution—we have supposed \( \ell = 1 \), equivalent to recasting the equation in dimensionless form with \( \ell \) chosen as the unit of length). Using \texttt{lsode}, we would invoke the statements

\[
\begin{align*}
& \text{>> b=1.5;} \\
& \text{>> axis equal;} \\
& \text{>> set( gca, 'xlim', [-4.0,4.0], 'ylim', [-4.0,4.0], 'fontsize', 14 );} \\
& \text{>> title( 'Open Orbits', 'fontsize', 16 )} \\
& \text{>> xlabel( 'X-Position', 'fontsize', 14 )} \\
& \text{>> ylabel( 'Y-Position', 'fontsize', 14 )} \\
& \text{>> hold on} \\
& \text{>> y0 = [-4.0 -3.0 -2.0 -1.0 1.0 2.0 3.0 4.0];} \\
& \text{>> for i= 1:8} \\
& \quad \text{ic=[-4.0; 1.0; y0(i); 0.0];} \\
& \quad \text{t = [0.0 : 0.25 : 60.0 ];} \\
& \quad [x, istate, msg] = lsode( @planet, ic, t ); \\
& \quad \text{plot( x(:,1), x(:,3), 'linewidth', 2.0, 'color', 'black' );} \\
& \text{end} \\
& \text{>> plot( [-4.0 0.0], [0.0 0.0], 'linewidth', 2.0, 'color', 'black' );} \\
& \text{>> grid on;}
\end{align*}
\]
Figure 11.15: Open orbits of a satellite moving under the influence of a sun at the origin.

```matlab
>> global k
  Set parameter.
>> k = 1.0;
  Set initial conditions.
>> ic = [0.0; 1.0];
  Set desired x coordinates.
>> x = [0.0 : 0.02 : 1.0];
>> [ f, istate, msg ] = lsode( @stdwaves, ...
                           ic, x );
  Invoke solver.
>> plot( x, f(:,1), 'linewidth', 2.0, ...
       'color', 'black' );
  Plot solution.
>> set( gca, 'ylim', [-1.0, 1.0] );
  Specify range on vertical axis.
>> set( gca, 'fontsize', 16 );
  Set size of axis labels
>> grid on;
```

The result is the solid line in Fig. 11.16. We are accepting that the default tolerance (Section 11.9.9) is adequate for this and subsequent calculations.

Clearly, however, the solution we have obtained (and in which we now have reasonable confidence) fails utterly to return to the value zero at the upper end of the interval. Either the assumed derivative \( g(0) = 1.0 \) or the assumed value \( k = 1.0 \) is not appropriate. Because the equations are linear, however, changing \( g(0) \) will merely influence the scale of the resulting solution; a different value of the derivative at the beginning will never cause the non-zero value at the end to become zero. We quickly conclude that efforts to bring \( f(1) \) to the value zero stand a chance of succeeding only if we tamper with \( k \). Thus, we explore other values of \( k \) with the statements

```matlab
>> hold on;
>> k = 2.0;
>> [ f, istate, msg ] = lsode( @stdwaves, ic, x );
>> plot( x, f(:,1), 'linewidth', 2.0, 'color', 'black', 'linestyle', ':' );
>> k = 3.0;
>> [ f, istate, msg ] = lsode( @stdwaves, ic, x );
>> plot( x, f(:,1), 'linewidth', 2.0, 'color', 'black', 'linestyle', '--' );
```
Figure 11.16: Shooting method to solve for standing waves in a string. The upper solid, dotted, dashed, and lower solid curves correspond to \( k = 1.0, 2.0, 3.0, \) and 4.0, respectively.

```matlab
>> k = 4.0;
>> [ f, istate, msg ] = lsode( @stdwaves, ic, x );
>> plot( x, f(:,1), 'linewidth', 2.0, 'color', 'black', 'linestyle', '-' );
>> hold off
```

All four tentative solutions are displayed in Fig. 11.16. Evidently, somewhere between \( k = 3.0 \) and \( k = 4.0 \), the solution will assume the proper value at \( x = 1 \), and we are now in a position via manual trial and error to seek more refined estimates of that specific value of \( k \).

Having now concluded that we should be adjusting \( k \) as we seek acceptable solutions and having recognized that the value of the solution at \( x = 1 \) is critical, we imagine that a graph showing \( f(1) \) as a function of \( k \) might help us in deciding where we should look in an effort to find additional solutions. Suppose we were to seek solutions for values of \( k \) ranging from 0 to 20 (say) in steps of 0.2. We might then solve the problem for each \( k \) but save only the value \( f(1) \) for each solution. To generate that information, we would have to create a loop that migrated through the values of \( k \), calculated the solution at each value, and saved the value \( f(1) \). We might use the statements

```matlab
>> global k
>> kk = 0:0.2:100.0/5.0;
>> ic = [0.0; 1.0];
>> x = [0.0 : 0.02 : 1.0];
>> for i = 1:101
    k = kk(i);
    [f, istate, msg] = lsode( @stdwaves, ic, x );
    soln(i) = f(51);
end
>> plot( kk, soln, 'linewidth', 2.0, 'color', 'black' )
>> xlabel( 'k', 'fontsize', 14 )
```
11.9. SOLVING ODES NUMERICALLY WITH OCTAVE

Figure 11.17: Graph of $f(1)$ versus $k$ for standing waves in a string.

```
>> ylabel( 'f(1)', 'fontsize', 14)
>> set(gca, 'fontsize', 14)
>> grid on
```

Here, we

- Make sure $k$ is declared global.
- Establish in $kk$ values that will be passed to $k$, one at a time.
- Set the initial conditions.
- Set the desired $x$ coordinates.
- Execute a loop that (a) selects the next value of $k$, (b) invokes `lsode` to solve the equations, and (c) stores in `soln(i)` the solution at $x = 1$ for the current value of $k$.
- Plot the resulting values.

Note that the loop requests a fair bit of computation and may take awhile to execute.

The resulting graph is shown in Fig. 11.17. Each of the zeroes of the function shown in this graph corresponds to a value of $k$ at which an acceptable solution to the boundary value problem can be found. Further, an approximation to each solution can be read from this graph and taken as input for a more sophisticated search procedure (which, however, we shall not develop here; see Chapter 14). Even more, we might infer from the oscillatory nature of this graph that the problem actually has a very large number of distinct solutions, only the first six of which are identifiable in the range $0 \leq k \leq 20$. The first six, however, suggest (correctly) that the solutions are equally spaced in $k$.

11.9.8 The Quantum Harmonic Oscillator

The quantum harmonic oscillator described by Eqs. (11.60) and (11.61) provides a second example of a boundary value problem. This one, however, involves an infinite domain. As we suggested
Table 11.15: The OCTAVE function qmshm.m.

function derivs = qmshm( psi, y )
# QMSHM: returns the derivatives for the quantum harmonic oscillator
# QMSHM defines the dimensionless Schrödinger equation for the quantum
# simple harmonic oscillator, which is
# 2
# d psi 2
# ----- + (2 ee - y ) psi = 0
# dy
# Entries are psi and d psi/dy, and initial conditions will thus be
# [psi0, dpsi0]. The parameter ee represents a dimensionless energy
# and is passed to qmshm as a global variable.

global ee
derivs = [ psi(2); -(2*ee-y^2)*psi(1) ];
endfunction

earlier, we can generate a more tractible approach by recognizing that the character of the equation
compels solutions to have either even or odd parity. Thus, we can focus attention on only the
interval \( 0 \leq y < \infty \). To set up the problem, we must first regard the second-order equation as
a pair of first-order equations in which we associate \( \psi(1) \) with \( \psi \) and \( \psi(2) \) with \( d\psi/dy \). The
equations we must solve then become

\[
\frac{d}{dy}\psi(1) = \psi(2) ; \quad \frac{d}{dy}\psi(2) = -(2\epsilon - y^2)\psi(1) \tag{11.105}
\]

and a suitable M-file qmshm.m defining these equations for OCTAVE’s solvers is shown in Table 11.15.
With this file stored in the default directory with the name qmshm.m, we are ready to seek its
solutions.

We seek first solutions of even parity, i.e., solutions for which \( \psi(0) \neq 0 \) and \( d\psi(0)/dy = 0 \). With
those values, we can address this boundary value problem as if it were an initial value problem—
except that we must reject solutions that do not go to zero as \( y \to \infty \). Only \( \psi(0) \) and \( \epsilon \) are adjustable
as we seek to impose the boundary condition at infinity. Since the ODE we are solving is linear,
however, we recognize right away that tampering with \( \psi(0) \) will merely affect the scaling of the
solution and has no power to convert a solution that doesn’t go to zero at infinity into one that does
go to zero. The only parameter we need bother adjusting is \( \epsilon \). Let us, therefore, standardize by
setting \( \psi(0) = 1.0 \) and explore the dependence of the solution on \( \epsilon \). At the outset, we do not know
what to expect. We do, however, believe that, in a dimensionless casting, important quantities are
likely to have values on the order of one. Thus, we begin a search for acceptable solutions by setting
\( \epsilon = 1.0 \) and examining the solution in the interval \( 0.0 \leq y \leq 2.0 \). The statements

```octave
>> global ee
>> ic = [ 1.0; 0.0 ]; ee = 1.0;
>> y = [0.0 : 0.1 : 4.0 ];
>> [psi, istate, msg] = lsode( @qmshm, ic, y );
>> plot( y, psi(:,1) );
>> axis( [0.0 4.0 -2.0 2.0 ] );
```

\(^{46}\)We here use the symbol \( y \) for what we earlier called \( \tau \).
generate that solution and produce a graph which—see your own output—starts off sensibly by
decaying from $\psi(0) = 1.0$ towards $\psi(y) = 0.0$ as $y$ increases, but then crosses the value $\psi(y) = 0.0$
at about $y = 1.2$ and gives every indication of heading off towards $-\infty$. With $\epsilon = 1.2$, the divergence
happens more quickly but with $\epsilon = 0.8$, the solution takes longer to diverge. Evidently, lowering $\epsilon$
from 1.0 moves in the right direction. With this background, we are tempted to examine the behavior
with several values of $\epsilon$ by executing the statements listed in Table 11.16. which step $\epsilon$ through the
indicated values, plot the solution at each step, add a horizontal line at $\psi = 0.0$, and label each
graph. From the resulting display, which is shown in Fig. 11.18, we conclude that $\epsilon = 0.5$ yields a
solution that goes to zero at infinity and we therefore argue that one allowed energy corresponds to
the value $\epsilon = 0.5$. If we now allow $\epsilon$ to increase beyond the value 0.7, we find the solution diverges more and more
rapidly for a time, but then begins to turn back towards the axis. Identically the same coding as
illustrated above, but with the starting line

```matlab
>> e = [ 1.0 1.5 2.0 2.5 3.0 3.5 ];
```
to set the values of $\epsilon$ and the lines

```matlab
>> text( 2.1, -1.2, '1.0', 'fontsize', 16 );
>> text( 1.7, -1.7, '1.5', 'fontsize', 16 );
>> text( 2.7, -1.8, '2.0', 'fontsize', 16 );
>> text( 3.5, -0.3, '2.5', 'fontsize', 16 );
>> text( 2.7, -1.8, '2.0', 'fontsize', 16 );
>> text( 3.0, 1.2, '3.0', 'fontsize', 16 );
```
Figure 11.18: The solution to the Schrödinger equation for an even state of the quantum harmonic oscillator when $\epsilon$ has the indicated values. These graphs reveal the ground state (or state of lowest energy) at $\epsilon = 0.5$.

Figure 11.19: The solution to the Schrödinger equation for an even state of the quantum harmonic oscillator when $\epsilon$ has the indicated values. Evidently, the graph at $\epsilon = 2.5$ converges to zero at large $y$. As we shall confirm in the next figure, this state is actually the second excited state of the harmonic oscillator.
to place the labels appropriately, produces the graph shown in Fig. 11.19. From this graph, we conclude that $\epsilon = 2.5$ yields another solution that goes to zero at infinity, and we argue that another allowed energy corresponds to the value $\epsilon = 2.5$. We might even be tempted to speculate—correctly as it turns out—that additional allowed energies for *even* states will correspond to the values $\epsilon = 4.5, 6.5, 8.5, \ldots$.

We turn next to the odd states, for which we set $\psi(0) = 0.0$ and $d\psi(0)/dy = 1.0$ before exploring solutions for various values of $\epsilon$. Indeed, the previous paragraphs suggest that we might expect to find the lowest *odd* state corresponding to $\epsilon = 1.5$. To test that expectation, we set the initial conditions and energies with the statements

```
>> ic = [ 0.0; 1.0 ];
>> e = [ 1.3 1.4 1.5 1.6 1.7 ];
```

Then we execute the statements

```
>> q = size(e);
>> ee = e(1);
>> [psi, istate, msg] = lsode( @qmshm, ic, y);
>> plot( y, psi(:,1), 'linewidth', 2.0, 'color', 'black' );
>> axis( [0.0 4.0 -2.0 2.0 ] );
>> hold on
>> for i = 2:q(2)
    ee = e(i);
    [psi, istate, msg] = lsode( @qmshm, ic, y);
    plot( y, psi(:,1), 'linewidth', 2.0, 'color', 'black' );
end
>> plot( [0.0,4.0], [0.0, 0.0], 'color', 'black' );
```

to generate the graph in Fig. 11.20 and the statements

```
>> text( 2.7, 1.7, '1.3', 'fontsize', 16 );
>> text( 3.5, 1.5, '1.4', 'fontsize', 16 );
>> text( 3.5, 0.1, '1.5', 'fontsize', 16 );
>> text( 3.4, -1.0, '1.6', 'fontsize', 16 );
>> text( 2.8, -1.5, '1.7', 'fontsize', 16 );
>> set( gca, 'fontsize', 16 )
>> xlabel( 'y', 'fontsize', 16 )
>> ylabel( '\psi(y)', 'fontsize', 16 )
>> hold off;
```

to label the individual curves in that graph. That the graph for $\epsilon = 1.5$ behaves properly at infinity confirms our suspicion that the first *odd* state occurs at that value of the energy. We further suspect that additional odd states will be found at $\epsilon = 3.5, 5.5, 7.5, \ldots$.

Finally, in broad terms, we conclude that the ground state of the quantum harmonic oscillator has even parity; that acceptable solutions occur when $\epsilon = n + 1/2$ with $n = 0, 1, 2, 3, \ldots$; and that odd and even states occur alternately as $n$ increases, i.e., that the parity of the $n$-th state is $(-1)^n$. Finally, remembering that the physical energy $E$ and the dimensionless energy $\epsilon$ are related by $E = \epsilon \hbar \omega$, we infer that the energies of the quantum harmonic oscillator are given by

$$E_n = \left(n + \frac{1}{2}\right) \hbar \omega \quad (11.106)$$

though we have direct evidence for this (correct) conclusion only for the lowest three states.
11.9.9 Changing Lsode Options

When the default values of the options that control the detailed action of \texttt{lsode} are not appropriate to the task at hand, these options can be changed. The available options with the default values are

- **absolute tolerance** (atol), which specifies the maximum absolute error in the dependent variables. This property may be a scalar, in which case it applies to all dependent variables, or it may be a vector, in which case each component specifies the absolute tolerance for the corresponding dependent variable. The default is \(1.49012 \times 10^{-8}\).

- **relative tolerance** (rtol), which specifies the maximum relative error in the dependent variables. This property can only be a scalar, and it applies to all dependent variables. The default is \(1.49012 \times 10^{-8}\).

- **integration method**, which is a string specifying which of the available methods is to be used. Two options are provided, the distinction being whether the system of equations is stiff or non-stiff\(^\text{47}\). Specifically
  - \texttt{adams} or \texttt{non-stiff}—the two values for the option are synonyms—for solving equations that are not stiff, and
  - \texttt{bdf} (backward differential formula) or \texttt{stiff}—the two values for the option are again synonyms—for solving equations that are stiff, though this method will also work for non-stiff equations.

Method \texttt{non-stiff} requires no Jacobian matrix; method \texttt{stiff} will use a supplied Jacobian matrix (see the OCTAVE manuals) or, absent that Jacobian, will use finite difference methods to compute the Jacobian. The default method is \texttt{stiff}.

\(^{47}\)When the solution changes very rapidly over a short time interval but a solution is sought over a much longer time interval, the equation(s) is (are) said to be \texttt{stiff}. Not every system of equations falls into this category.
11.9. SOLVING ODES NUMERICALLY WITH OCTAVE

- **initial step size**, which specifies a starting value for the step size, which is then adjusted as the adaptive procedure works out the solution. The default is $-1$, which tells OCTAVE to make its own choice.

- **maximum order**, which limits how extensively the procedure will be refined as the solution is worked out. Except for the Adams method (for which the value can be between 1 and 12), this option must be between 1 and 5, inclusive. The default is $-1$, which tells OCTAVE to make its own choice.

- **maximum step size**, which specifies the largest step size the solver is to use. The default is $-1$, which tells OCTAVE to make its own choice.

- **minimum step size**, which, if specified, overrides the default value of 0.

- **step limit**, which, if specified, overrides the default maximum number—100,000—of steps.

The two tolerances are both taken into account as OCTAVE assesses convergence. Specifically, OCTAVE seeks to keep the absolute value of the local error in each of the returned values of the dependent variables no bigger than the quantity

$$ |\text{local error in } x(i)| \leq \text{rtol} \times |x(i)| + \text{atol}(i), \text{ all } i $$

where $x(i)$ is the returned value of the solution at every step along the way. Typically, when the solution is close to zero, the absolute tolerance provides the controlling criterion while, when the solution is well away from zero, the relative tolerance may dominate over the absolute tolerance. Either (but certainly not both) of these tolerances can be set to zero, in which case the other provides the controlling criterion throughout the solution. Be warned, however: if the absolute tolerance is set to zero, convergence may be difficult when the solution itself is close to zero (since a small fraction of zero—relative tolerance—may result in excessive computation).

If we choose to establish values different from the defaults for any of these options, we would do so with the command `lsode_options`, e.g.,

```
lsode_options( 'integration method', 'adams' )
```

Apparently, only one option can be modified by this command, so modification of several options will require that many separate comments. The statement `lsode_options()` will display the current values of these options.

To this point, we have accepted the default tolerances in each of the sample solutions. If, for example, in the solution of the problem in radioactive decay, we had wanted to change the absolute tolerance to 5.0 and the relative tolerance to 10.0%, we would have used the statements

```matlab
>> lsode_options( 'absolute tolerance', 5.0 ) Set tolerances.
>> lsode_options( 'relative tolerance', 0.1 )
>> global kA kB Declare global variables
>> kA = 0.1; kB = 0.1;
>> t = [0.0 : 0.25 : 50.0]; Set parameters.
>> ic = [1000.0; 0.0; 0.0 ]; Set desired times for solution.
>> [n, istate, msg] = lsode( @decaymod, ... Set initial conditions.
   ic, t ); Solve the system with the declared
>> plot( t, n ) tolerances.
```

Confirmation that the solution generated by this coding and the solution generated with the default tolerances differ by very little is left as an exercise for the reader.

Further detail about the options that affect the behavior of `lsode` can be found in the OCTAVE manuals.
11.9.10 M-files in the Public Library

The following M-files are found in the directory $HEAD/octave (and may also at some sites have been placed in a directory in OCTAVE’s default search path) and can be used to explore the systems they describe. Further information about the file decay.m, for example, can be obtained either with the command

>> help, ' $HEAD/octave/decay'

to OCTAVE or with whatever command in your operating system lists the file on the screen. The first of these commands instructs OCTAVE to find the file decay.m and display the comments in the file starting with the second line and continuing until the first executable statement is reached; the second command, which must be executed in a command window to the operating system (not within OCTAVE), will display the entire file on the screen. Many of the functions defined by these files involve parameters that must be supplied as global variables; examine the help message and/or the file itself to determine the proper specification of these variables.

- apollo.m: describes a satellite in the gravitational attraction of two fixed suns of equal mass.
- dampharm.m: describes the one-dimensional damped, unforced harmonic oscillator using dimensional variables.
- damposc.m: describes the one-dimensional, damped, unforced harmonic oscillator using dimensionless variables.
- decay.m: describes three-species radioactive decay.
- decaymod.m: describes three-species radioactive decay using global variables to communicate parameters to the function.
- drvnosc.m: describes the one-dimensional, damped, forced harmonic oscillator using dimensional variables.
- drvnpend.m: describes the forced, damped simple pendulum in dimensionless units.
- henon.m: describes the Hénon-Heiles oscillator.
- largamp.m: describes the undamped, unforced large amplitude pendulum in dimensionless units.
- lorenz.m: describes the Lorenz attractor.
- onedshm.m: describes the one-dimensional, undamped, unforced harmonic oscillator using dimensional variables.
- planet.m: describes the motion of a planet of mass $m$ around a sun of mass $M$, using dimensionless units. An additional parameter allows for exploration of non-inverse square forces.
- qmshm.m: describes the dimensionless quantum harmonic oscillator.
- rossler.m: describes the Rössler attractor.
- stdwaves.m, which describes the behavior of standing waves in a string.
- twodshm.m: describes the two-dimensional, undamped, unforced harmonic oscillator using dimensionless variables.

48 In UNIX and windows, one possible command would be more $HEAD/octave/decay.m; see the Local Guide.
11.13. Solving ODEs Numerically with Mathematica

Mathematica’s command \texttt{NDSolve} provides a means to solve ordinary differential equations numerically. A full statement using \texttt{NDSolve} will have the form

\[
\text{NDSolve} \left\{ \begin{array}{l}
\text{ODE}_1, \text{ODE}_2, \ldots, \text{IC}_1, \text{IC}_2, \ldots,
\{ \text{DVar}_1, \text{DVar}_2, \ldots \},
\{ \text{IVar}, \text{IVarMin}, \text{IVarMax} \} \right\}
\]

where \text{ODE} and \text{IC} are the differential equations and initial conditions, \text{DVar} represents a dependent variable, and \text{IVar}, \text{IVarMin}, and \text{IVarMax} are the independent variable and the lower and upper limits on the range in which the solution is to be found. In this case, both the differential equations and the initial conditions must be specified; Mathematica will not introduce undetermined integration constants. Further, all parameters in the equations must be given explicit numerical values before \texttt{NDSolve} is executed. When invoked, \texttt{NDSolve} returns not a solution but an interpolation function that, once a specific numerical value has been assigned to the independent variable, can be used to generate a numerical value for the solution at the specified value of the independent variable. Thus, for example, to address the problem of chain radioactive decay that we have already solved symbolically in Section 11.5.4, we would request Mathematica to generate the interpolation function by invoking the statements

\begin{verbatim}
In[1]:= eqA = D[ AA[t], t ] == -kA*AA[t];
In[2]:= eqB = D[ BB[t], t ] == kA*AA[t] - kB*BB[t];
In[3]:= eqC = D[ CC[t], t ] == kB*BB[t];
In[4]:= kA = 0.1; kB = 0.1;
In[5]:= soln = NDSolve[ {eqA, eqB, eqC, AA[0]==1000, BB[0]==0, CC[0]==0}, {AA, BB, CC}, {t, 0, 50} ]
Out[5]= {{AA -> InterpolatingFunction[Graphic], BB -> InterpolatingFunction[Graphic],
     CC -> InterpolatingFunction[Graphic]}}
\end{verbatim}

Then, after the interpolating function has been created, the solution at any particular point can be found with statements like

\begin{verbatim}
In[6]:= { AA[0.0], BB[0.0], CC[0.0] } /. soln
Out[6]= {{1000., 0., 0.}}
In[7]:= { AA[0.1], BB[0.1], CC[0.1] } /. soln
Out[7]= {{990.05, 9.9005, 0.0496679}}
\end{verbatim}

\footnote{Because \texttt{C} is a protected variable in Mathematica, the dependent variables here are named \texttt{AA}, \texttt{BB}, and \texttt{CC} rather than \texttt{A}, \texttt{B}, and \texttt{C}.}

\footnote{The inclusion identified here with the word ‘Graphic’ is a box that has a small graph of the function as well as an indication of the domain of the solution and the character of the output. Clicking \texttt{ML} on the ‘+’ sign in that graphic brings up information about other features of the solution.}
Note the form of the solution, which is presented as a list of lists, though this particular solution has only one list in the outer list. In that (one) list, the elements are the numerical values of the solution at the selected value of the independent variable—time in this case.

Should we wish a graph of the solution over some range of the independent variable, we would invoke the command `Plot` but use the command `Evaluate` to determine the values to be plotted. A graph similar to Fig. 11.9 for this decay would, for example, be displayed by the statement

```mathematica
In[8]:= Plot[ Evaluate[ {AA[t], BB[t], CC[t]} /. soln, {t, 0.0, 50.0} ],
        PlotStyle -> Thickness[0.01], AxesLabel -> {"t", "A,B,C"} ]
```

Additional keywords similar to those available for the command `Plot` can be exploited to add elements to the graph and/or override additional defaults.

The previous two paragraphs illustrate the essential strategy adopted by `Mathematica` to produce numerical solutions to ordinary differential equations. We here draw attention to a few additional capabilities, leaving an exhaustive enumeration and the details to the `Mathematica` manuals. While the defaults adopted for controlling accuracy and for protecting against infinite loops are often acceptable, the command `NDSolve` nonetheless possesses several options that give us the ability to override these defaults. Among these options are

- **WorkingPrecision** (default `$MachinePrecision`, often 16), which specifies the number of digits to be used for internal computations.
- **AccuracyGoal** (default usually figured as half the value of `WorkingPrecision`), which specifies the absolute accuracy to be sought by stipulating the number of digits to be provided in the solution. `AccuracyGoal` should not exceed `WorkingPrecision`.
- **PrecisionGoal** (default usually figured as half the value of `WorkingPrecision`), which specifies the relative accuracy sought in the final result. `PrecisionGoal` should not exceed `WorkingPrecision`.

*Note:* With `AccuracyGoal` set to $a$ and `PrecisionGoal` set to $p$, `Mathematica` strives for an absolute accuracy in a value $x$ to be less than $10^{-a} + 10^{-p} |x|$.

- **MaxSteps** (default automatic), which specifies the maximum number of steps to take in generating the solution. This option provides a means to stop the iterations even if goals on accuracy and precision have not been met.
- **MaxStepSize** (default automatic), which limits the size of the largest steps to be taken in generating the solution. This option provides a means to specify, for example, a maximum step size smaller than `Mathematica` determines to be adequate.

Remember, too, that the statement `Options[NDSolve]` will display all options and their current default values for the indicated command.

If we want solutions at a specific set of values of the independent variable, we could exploit the command `Table`. For example, the statement

```mathematica
In[9]:= tbl = Table[ {t, AA[t], BB[t], CC[t]} /. soln, {t, 0, 20, 5} ]
```

produces a table named `tbl` whose $i$-th element contains four values (the time and the values of $A$, $B$, and $C$ at that time). This output can then be manipulated using any of `Mathematica`'s resources for working with tables.
11.16  Exercises

11.16.1  . . . using Symbolic Methods

11.1.  Find the motion of a driven, damped oscillator satisfying the differential equation

\[ \frac{d^2 x}{dt^2} + \frac{b}{m} \frac{dx}{dt} + \frac{k}{m} x = m f \cos \omega t \]

subject to the general initial conditions

\[ x(0) = x_0 \quad ; \quad \frac{dx}{dt}(0) = v_0 \]

and then determine the particular initial values that might be imposed so that the transient part of the solution is wiped out from the beginning, i.e., so that the motion is identically the steady-state motion from the moment the oscillator is set into motion. Assume that the oscillator is underdamped. (Note that, so that \( m \) will ultimately appear only in conjunction with \( b \) and \( k \), we have chosen to define \( f \) as a force per unit mass rather than simply a force.)

11.2.  Take away the walls and the two springs connecting the blocks to the walls in the system of Fig. 11.5, let the masses be different \((m_1 \text{ and } m_2, \text{ say})\) and denote the constant of the one spring by \( k \). Suppose the blocks are constrained to move along a straight line. Measuring from an arbitrarily selected origin on that line, let the coordinates of the particles be \( x_1 \) and \( x_2 \), respectively. The equations of motion for this system are

\[ m_1 \frac{d^2 x_1}{dt^2} = k (x_2 - x_1) \quad ; \quad m_2 \frac{d^2 x_2}{dt^2} = -k (x_2 - x_1) \]

Let the system be put into motion with arbitrary initial conditions

\[ x_1(0) = x_{10} \quad ; \quad x_2(0) = x_{20} \quad ; \quad \frac{dx_1(0)}{dt} = v_{10} \quad ; \quad \frac{dx_2(0)}{dt} = v_{20} \]

Solve this initial-value problem for \( x_1(t) \) and \( x_2(t) \) and then examine the behavior of the particular quantities

\[ X(t) = \frac{m_1 x_1(t) + m_2 x_2(t)}{m_1 + m_2} \quad \text{and} \quad Y(t) = x_2(t) - x_1(t) \]

which are, respectively, the position of the center of mass of the system and the position of the second block relative to the first block.

11.3.  Study the behavior of the system that results when the system shown in Fig. 11.5 is extended to contain three objects coupled in a line. Take the four springs all to have the same spring constant but allow for the possibility that the middle object may have a mass different from that of the two outside objects. Your examination should include at least the following:

- Find both the natural frequencies of the normal modes of oscillation and the initial conditions that will excite the system exclusively in each of the normal modes, determining particularly the frequencies as a function of the ratio of the two masses. To display your results, draw a diagram something like a quantum energy level diagram that shows the way the three natural frequencies vary as the middle mass goes from being rather smaller than the two outer masses to being rather larger than the outer two masses. As initial conditions, it will be sufficient to release the objects from rest at individually arbitrary displacements. You are, however, quite likely to have to help with the inverse Laplace transforms, since your symbolic manipulator may be unable to deal with the inverse Laplace transform of an expression like

\[ \frac{A s + B s^3}{\alpha s^4 + \beta s^2 + \gamma} = s \left( \frac{A + B s^2}{\alpha s^4 + \beta s^2 + \gamma} \right) \]

The denominator looks quartic in \( s \) but is better seen as quadratic in \( s^2 \). If you introduce \( z = s^2 \) temporarily, factor the resulting denominator, and then do a partial fraction expansion,
you can end up with the alternative form
\[
s \left( \frac{A + B s^2}{\alpha s^4 + \beta s^2 + \gamma} \right) = s \left( \frac{R}{s^2 + R'} + \frac{Q}{s^2 + Q'} \right) = \frac{Rs}{s^2 + R'} + \frac{Qs}{s^2 + Q'}
\]
whose inverse Laplace transform is easier to find. (Indeed, even Table 1.2 is adequate to the task.) Finding the constants \( R, A, R' \) and \( Q' \) is part of the exercise; the expressions are not pretty. The lesson: Symbolic manipulators frequently need help from someone who really knows where the calculation is going.

- With the three masses equal, find the solution for \( x_1(t) \), \( x_2(t) \), and \( x_3(t) \) when one of the outer masses is initially displaced, the other two are not, and all three are released from rest. Plot graphs of all three positions as functions of time over a long enough time interval to reveal the features of the motion.

- With the three masses equal, find the solution for \( x_1(t) \), \( x_2(t) \), and \( x_3(t) \) when the middle mass is initially displaced, the other two are not, and all three are released from rest. Plot graphs of all three positions as functions of time over a long enough time interval to reveal the features of the motion.

To help you get started and to facilitate focusing on the solution of the ODEs rather than on deriving them, note that, for three masses, the equations of motion will be
\[
\begin{align*}
m \frac{d^2 x_1}{dt^2} &= -kx_1 + k(x_2 - x_1) \\
m' \frac{d^2 x_2}{dt^2} &= -k(x_2 - x_1) + k(x_3 - x_2) \\
m'' \frac{d^2 x_3}{dt^2} &= -k(x_3 - x_2) - kx_3
\end{align*}
\]

11.4. The system called the double pendulum shown in Fig. 11.21 consists of a ball of mass \( m_1 \) hanging from a rigid and massless rod of length \( l_1 \) attached to the ceiling and a second ball of mass \( m_2 \) hanging from a rigid and massless rod of length \( l_2 \) attached to the first ball. The balls swing in a plane, and the configuration of the system is specified by giving two angles, the first of which, \( \theta \), gives the angle that the upper string makes with the vertical and the second of which, \( \phi \), gives the angle that the lower string makes with the vertical. The motion can be very complicated and at times will be chaotic. For small amplitudes, however, things are much more sedate. When the amplitudes of the motion of both balls are small and—to simplify a little bit—when the strings are both the same length (\( l_1 = l_2 \), which we will symbolize with the letter \( l \)), the equations of motion turn out to be
\[
\begin{align*}
\frac{d^2 \theta}{dt^2} + \frac{m_2}{m_1 + m_2} \frac{d^2 \phi}{dt^2} + \frac{q}{l} \theta &= 0 \\
\frac{d^2 \phi}{dt^2} + \frac{d^2 \theta}{dt^2} + \frac{q}{l} \phi &= 0
\end{align*}
\]
Find the normal modes of oscillation of this system and determine the initial conditions that will cause the system to oscillate exclusively in one or the other of these modes.
11.5. Solve the differential equation
\[ m \frac{d^2 x}{dt^2} + b \frac{dx}{dt} + kx = mf \cos \omega t \]
for the driven, damped harmonic oscillator subject to the specific initial conditions
\[ x(0) = x_0 \; ; \; \frac{dx}{dt}(0) = 0 \]
by explicitly taking the Laplace transform of the equation, solving that result for the Laplace transform of the solution, and then inverting that transform to obtain the solution itself.

11.6. Among the simplest of differential equations is the equation
\[ \frac{d^2 x}{dt^2} + \omega^2 x = 0 \]
that describes a simple harmonic oscillator. Generate a series solution to this equation and then verify that the solution thus generated agrees with the known solution
\[ x(t) = A \cos \omega t + B \sin \omega t \]
where \( A \) and \( B \) are constants determined by the initial conditions.

11.7. Find a symbolic solution for all three components for the motion of a projectile in a linear, viscous medium, when the initial conditions are general, i.e., solve Eq. (11.4) subject to the initial conditions in Eq. (11.2). Since the equations are uncoupled, you can solve each individually. Alternatively, you can solve the three equations simultaneously as a system. Solve them both ways. Once you have the solutions in hand, verify that they satisfy the original equations and initial conditions. Finally, explore their limits for small \( b \).

11.8. Recast Eq. (11.34) for the LRC circuit in dimensionless form, measuring charge in units of \( q_0 \) and finding a suitable unit in terms of which to measure time.

11.9. Reflecting Coulomb’s law, the equation of motion for a particle carrying charge \( q \) moving in the (fixed) electrostatic field of a charge \( Q \) is
\[ m \frac{d^2 r}{dt^2} = \frac{qQ}{4\pi\varepsilon_0} \frac{r}{r^3} \]
Here, the force center is assumed to be at the origin, \( r \) is the position vector of the particle, and the equation is written in mksa units. Write this equation of motion out in Cartesian coordinates, cast it in dimensionless form, and determine the correspondences that must be adopted to convert the equations for an orbit in an electrostatic field into those for the orbit in a gravitational field, i.e., into Eq. (11.46).

11.10. Recast the system of six first-order equations in the last paragraph of Section 11.1.1 in dimensionless form, both when the viscous term is linear in the velocity and when it is quadratic in the velocity. Appropriate units for this recasting can be inferred from the discussion earlier in that section.

11.11. Translate the entire discussion—both dimensional and dimensionless—of the planetary problem in Section 11.1.7 from Cartesian coordinates \((x, y)\) to polar coordinates \((r, \phi)\), where \( x = r \cos \phi \) and \( y = r \sin \phi \). That is
(a) Show that the polar components of the equations of motion in the first instance are
\[ m \left[ \frac{d^2 r}{dt^2} - r \left( \frac{d\phi}{dt} \right)^2 \right] = f(r) ; \quad m \left( \frac{d^2 \phi}{dt^2} + 2 \frac{dr}{dt} \frac{d\phi}{dt} \right) = 0 \]
(b) Recognizing that
\[ \frac{d}{dt} \left( mr^2 \frac{d\phi}{dt} \right) = m \left( r^2 \frac{d^2 \phi}{dt^2} + 2 \frac{dr}{dt} \frac{d\phi}{dt} \right) \]
show that
\[ mr^2 \frac{d\phi}{dt} = \text{constant} = L \implies \frac{d\phi}{dt} = \frac{L}{mr^2} \]
and then that
\[ \frac{d^2 r}{dt^2} = f(r) + \frac{L^2}{mr^3} \]
Here, \( L = m(x_0v_{y0} - y_0v_{x0}) \) is the angular momentum of the object; \( L \) is constant throughout the motion. (In polar coordinates, we would first solve this single, second-order, non-linear, inhomogeneous equation for \( r(t) \). Then, with \( r(t) \) in hand, we integrate the equation \( \frac{d\phi}{dt} = \frac{L}{mr^2} \)—see Chapter 13—to find \( \phi(t) \).

(c) Translate the initial conditions to polar coordinates, finding that
\[ r(0) = \sqrt{x_0^2 + y_0^2} \]\n\[ \frac{dr}{dt}(0) = \frac{x_0v_{x0} + y_0v_{y0}}{\sqrt{x_0^2 + y_0^2}} \]\n\[ \phi(0) = \tan \frac{y_0}{x_0} \]\n\[ \frac{d\phi}{dt}(0) = \frac{x_0v_{y0} - y_0v_{x0}}{x_0^2 + y_0^2} \]

(d) Restrict the force to the inverse square gravitational force, finding that
\[ \frac{d^2 r}{dt^2} = -\frac{GM}{r^2} + \frac{L^2}{mr^3} \]

(e) Recast the differential equations and initial conditions in dimensionless form, finding that
\[ \frac{d^2 \tau}{dt^2} = -\frac{1}{\tau^2} + \frac{\beta}{\tau^3} \]\n\[ \frac{d\phi}{dt} = \pm \sqrt{\beta} \]
\[ r(0) = \frac{r_0}{\ell} = \frac{\sqrt{x_0^2 + y_0^2}}{\ell} \]\n\[ \phi(0) = \tan \frac{y_0}{x_0} \]\n\[ \frac{dr}{dt}(0) = \frac{x_0v_{x0} + y_0v_{y0}}{\sqrt{x_0^2 + y_0^2} \sqrt{GM/\ell}} \]\n\[ \frac{d\phi}{dt}(0) = \frac{x_0v_{y0} - y_0v_{x0}}{(x_0^2 + y_0^2) \sqrt{GM/\ell^3}} \]
where \( \beta = \frac{L^2}{GMm^2\ell} \) is a constant, and—in the second equation—the upper sign applies when \( L > 0 \) and the lower sign when \( L < 0 \). While all parameters disappeared from the equations in Cartesian coordinates, the parameter \( \beta \) remains in the equations in polar coordinates.

(f) Recast the Cartesian expressions in the the last paragraph of Section 11.1.7 for both the dimensional and the dimensionless statements of conservation of energy and angular momentum in the planetary problem into polar coordinates.

**Hint:** Remember (or take as given if you haven’t met them yet) that the radial and azimuthal components of the acceleration of a particle in polar coordinates are given by \( a_r = \ddot{r} - r\dot{\phi}^2 \) and \( a_\phi = r\ddot{\phi} + 2\dot{r}\dot{\phi} \).

### 11.16.2 ... using Numerical Methods

#### 11.12.
In an appropriate dimensionless presentation, standing waves in a string must satisfy the boundary value problem
\[ \frac{d^2 y}{dx^2} + k^2 y = 0 \quad ; \quad y(0) = y(1) = 0 \]
Suppose that the interval \( 0 \leq x \leq 1 \) is divided into \( n \) equal segments of length \( \Delta x = 1/n \), let \( x_i = i\Delta x \) (with \( i = 0, 1, 2, \ldots, n \)), and let \( y_i = y(x_i) \). Evaluate the ODE at \( x = x_i \), approximate the second derivative with the difference formula
\[ \frac{d^2 y}{dx^2} \bigg|_{x=x_i} \approx \frac{y_{i+1} - 2y_i + y_{i-1}}{\Delta x^2} \quad ; \quad i = 1, 2, 3, \ldots n - 1 \]
and note that \( y_0 = y_n = 0 \). Show that the values \( y_i \) for \( i = 0, 1, 2, \ldots, n \) satisfy a system of \( n + 1 \) linear algebraic equations of the form \( MY = \alpha Y \), where \( Y \) is an \((n + 1)\)-component vector whose elements are the values of \( y_i \), and \( \alpha \) is determined from \( k^2 \) and \( \Delta x \). Then argue that the allowed values of \( k^2 \) can be determined from the eigenvalues of the matrix \( M \). That is, show that this transformation turns a boundary value problem involving a differential equation into an approximately equivalent matrix eigenvalue problem.

11.14. Some radioisotopes exhibit a branched decay sequence such as

\[
\begin{array}{c}
A \quad k_1 \quad B \\
\quad k_2 \quad C \quad k_3 \quad D \\
\end{array}
\]

Deduce appropriate differential equations, create an appropriate file to define the equations, and thoroughly explore the behavior of this system. Assume that \( D \) is stable and that, initially, only \( A \) is present.

11.15. Taking the initial conditions for the unforced, damped harmonic oscillator of Eq. (11.27) with \( \vec{F} = 0 \) to be \( \vec{v}(0) = 1 \) and \( v(0) = 0 \), produce graphs of \( \vec{v} \) versus \( \vec{t} \) for several values of \( \beta \) on the interval \( 0.0 < \beta < 2.0 \). Then using the knowledge that \( \beta = b/m\omega = b/\sqrt{mk} \), and that \( \vec{t} = \omega \vec{t} \), infer from your graphs the effect of changing \( m \) or \( k \) on the physical motion.

11.16. Explore the unforced, damped harmonic oscillator of Eq. (11.27) with \( \vec{F} = 0 \) for the cases of critical damping (\( \beta = 2.0 \)) and overdamping (\( \beta > 2.0 \)).

11.17. Explore the behavior of the Van der Pol oscillator described in dimensionless form by the equation

\[
\frac{d^2\vec{r}}{dt^2} = \frac{d\vec{r}}{dt}(1 - \vec{r}^2) - \vec{r}
\]

obtaining graphs of position versus time, velocity versus time, and velocity versus position (the phase-plane trajectory), each for several different initial conditions. Convince yourself that the final, steady-state path in the phase plane is independent of the initial conditions.

11.18. The angular position \( \theta \) of a simple pendulum of length \( l \) satisfies the non-linear equation

\[
\frac{d^2\theta}{dt^2} + \frac{g}{l} \sin \theta = 0
\]

where \( \theta \) is measured in radians from the lowest point of the pendulum’s motion. Use numerical methods to study the motion of this pendulum when it is released from rest at each of several initial displacements, say 20°, 45°, 90°, 120°, 150°, 165°, and 178°. Look particularly at graphs of \( \theta \) versus \( t \), \( d\theta/dt \) versus \( t \), and \( d\theta/dt \) versus \( \theta \) (the phase plot). Obtain also a graph of period versus amplitude (initial displacement). Write several paragraphs describing your set up of the problem and presenting evidence for your discoveries. Optional: Try starting the pendulum at the bottom (0 initial angle) with several initial angular velocities. How large can the angular velocity be before the pendulum begins to swing over the top? Suggestion: Begin by introducing the dimensionless time \( \vec{t} = \sqrt{g/l} \) so that the equation becomes \( d^2\theta/d\vec{t}^2 + \sin \theta = 0 \).

11.19. Suppose that the “gravitational” force were not inverse square but instead depended on some other (negative) power of the radial coordinate. The dimensionless equations of motion then would be

\[
\frac{d^2\vec{r}}{d\vec{t}^2} = -\frac{\vec{r}}{(\vec{r}^2 + \vec{y}^2)^{\frac{3}{2}}} ; \quad \frac{d^2\vec{y}}{d\vec{t}^2} = -\frac{\vec{y}}{(\vec{r}^2 + \vec{y}^2)^{\frac{3}{2}}}
\]

[Compare Eq. (11.46).] Of course, the equations reduce to those for the inverse square force if we simply set \( b = 3/2 \). For the planetary problem, find conditions that will generate a distinctly elliptical orbit for an attractive inverse square force (\( b = 1.5 \)). Then explore the effect on that orbit of distorting the force by changing the exponent in the denominator of the equations making
Suppose a particle of charge $q$ and mass $m$ is injected into a region of space containing constant, crossed electric and magnetic fields $\mathbf{E} = E_i \hat{i}$ and $\mathbf{B} = B_k \hat{k}$. In vector form, the equation of motion for this particle is
\[
m \frac{d^2 \mathbf{r}}{dt^2} = q \mathbf{E} + q \frac{d}{dt} \times \mathbf{B}
\]
Verify the equations of motion
\[
m \frac{d^2 x}{dt^2} = q E_x + q B_z \frac{dy}{dt} ; \quad m \frac{d^2 y}{dt^2} = -q B_z \frac{dx}{dt} ; \quad m \frac{d^2 z}{dt^2} = 0
\]
for the specific fields of this exercise, express them in dimensionless form (note that $\omega = q B_z / m$ is a frequency and $E_x / B_z$ is a velocity), and thoroughly explore the behavior of the particle in this situation. Try to understand the motion intuitively. Hint: You should find that, in terms of an arbitrarily selected unit of length $\ell$, the equations involve a single parameter $q E_x / (m \omega^2 \ell)$, which can alternatively be written as $(E_x / B_z) / (\omega \ell)$ — the ratio of the velocity $E_x / B_z$ determined by the fields to the characteristic velocity implied by your choice of a length unit and the frequency $\omega$. Note that this exercise actually has more than one parameter, since the initial components of the velocity—probably expressed in units of $\omega \ell$—also influence the solution.

A particle having mass $m$ and carrying charge $q$ moves in the $xy$-plane while experiencing an electric field given by $\mathbf{E}(x, y) = -\alpha y \hat{j}$, where $\alpha$ is a constant. Assume that $\alpha$ and $q$ are both positive. With $\mathbf{r} = x \hat{i} + y \hat{j}$, the vector equation of motion for this particle then is
\[
m \frac{d^2 \mathbf{r}}{dt^2} = -q \alpha y \hat{j}
\]
(a) Show that, in component form, the equations of motion for this particle are
\[
\frac{d^2 x}{dt^2} = 0 ; \quad \frac{d^2 y}{dt^2} = -\alpha q \frac{y}{m} = -by
\]
(b) With $b$ a global parameter, create an appropriate file to define these equations for the ODE-solver in an available numerical/graphical tool. (c) Use that tool to obtain graphs of the trajectories in the $xy$-plane of several particles projected from the origin at different angles and with different speeds. (d) Speculate on a use for this field.

An important system in the early study of chaos is described by the Lorenz equations
\[
\begin{align*}
\frac{dx}{dt} &= a(y - x) \\
\frac{dy}{dt} &= -xz + bx - y \\
\frac{dz}{dt} &= xy - cz
\end{align*}
\]
Create an appropriate file defining these equations and then thoroughly explore the behavior of this system. Graphs of $y$ versus $x$, $z$ versus $x$, and $z$ versus $y$ when $a = 10.0$, $b = 28.0$ and $c = 8.0/3.0$. 

11.20. Explore the scattering orbits that occur when an object moves under the action of a repulsive inverse square force, and compare your results with those for an attractive force.

11.21. Deduce the equations of motion for a space ship of mass $m$ coasting freely in the $xy$ plane under the gravitational influence of two suns, each of mass $M$ and located respectively at $(R, 0)$ and $(-R, 0)$. Then express the equations in dimensionless form and, creating all necessary files, thoroughly explore the motion of this space ship. In particular, you might search for an orbit that loops like a figure-eight around the two suns and/or you might see if your approach predicts what you would expect intuitively if you start the spaceship from rest at a point on the perpendicular bisector of the line joining the two suns. Make sure your solutions are generated to an adequate accuracy.
under the initial conditions \( x_0 = 1.0, y_0 = 0.0, \) and \( z_0 = 0.0 \) are particularly interesting. While graphs of \( y \) versus \( x \), \( z \) versus \( x \), and \( z \) versus \( y \) are interesting, the true beauty of the trajectory is best seen using a three-dimensional space curve. Be sure to examine the path from several different vantage points in the space around the path, an objective most easily accomplished if the display of the path allows rotation of the path on the screen.

11.27. The dynamics of the chemical reaction

\[ A + B \leftrightarrow C + D \]

is governed by the equations

\[ \frac{dA}{dt} = -k_f AB + k_r CD \quad \frac{dB}{dt} = -k_f AB + k_r CD \]

\[ \frac{dC}{dt} = k_f AB - k_r CD \quad \frac{dD}{dt} = k_f AB - k_r CD \]

where \( A(t) \), \( B(t) \), \( C(t) \), and \( D(t) \) are the concentrations of each molecule in the reaction vessel, and \( k_f \) and \( k_r \) are the forward and reverse rate constants, respectively. Suppose that the reaction is started with \( A(0) = A_0 \), \( B(0) = B_0 \), and \( C(0) = D(0) = 0 \). Cast the equations in dimensionless form, using \( A_0 \) as the unit of concentration and \( k_f A_0 t \) as the dimensionless time. Then explore the behavior of the system as a function of the initial concentration of \( B \), measured in units of \( A_0 \) and the reverse rate constant, measured in units of \( k_f \). Look particularly at the dependence of the ultimate equilibrium on these parameters. Make sure your results are generated to adequate accuracy.

11.28. In classical ecology, the interaction between a predator and a prey, with populations \( x(t) \) and \( y(t) \), respectively, is modeled with the equations

\[ \frac{dx}{dt} = -k_1 x + k_2 xy \quad \frac{dy}{dt} = -k_3 y - k_4 xy \]

where \( k_1 \) and \( k_3 \) are parameters describing the way each population would evolve in the absence of the other and \( k_2 \) and \( k_4 \) are parameters describing strength of the interaction between the two species, which we take to be proportional to the likelihood of an encounter between a member of one species and a member of the other species. Depending on the parameters and the initial populations \( x(0) = x_0 \), \( y(0) = y_0 \), the system may approach a stable equilibrium or, alternatively, one or the other of the populations may become extinct. Explore this system to determine conditions under which each of these circumstances occurs and write a paragraph or two describing your findings. Make sure your results are generated to an accuracy adequate to support your conclusions.

11.29. Examine the undamped, unforced harmonic oscillator carefully, using at least two different methods and several time steps. Monitor the accuracy of your solution by monitoring the constancy of the energy of the oscillator given (in dimensional form) by \( E = \frac{1}{2}mv^2 + \frac{1}{2}kx^2 \).

11.30. In the text, we illustrated the use of one or more numeric processing programs to solve the chain radioactive decay problem of Section 11.1.2 using Euler’s method. Using an available numeric processing program like IDL, MATLAB, OCTAVE, or PYTHON, repeat that development using the improved Euler method and comment on what the resulting graphs reveal about the adequacy of the several time steps used.

11.31. The type of child’s swing shown in Fig. 11.22 is hung with elastic ropes. Suppose the ropes are long enough so that the child/swing can be represented by a point mass at the end of a spring, assume the spring obeys Hooke’s law with constant \( k \) and has an unstretched length \( a \), and let the motion of the mass be confined to a single vertical plane. Show that, in the coordinate system illustrated, the equations of motion are

\[ m \frac{d^2x}{dt^2} = -kx + \frac{kax}{\sqrt{x^2 + y^2}} \quad m \frac{d^2y}{dt^2} = mg - ky + \frac{kay}{\sqrt{x^2 + y^2}} \]
Then, introducing $\omega_0^2 = k/m$, $\bar{t} = \omega_0 t$, $\bar{x} = x/a$, and $\bar{y} = y/a$, cast the equations in dimensionless form. After the suggested rescalings, only one parameter $g/a\omega_0^2$—remains and this parameter is the square of the ratio of the frequency $\sqrt{g/a}$ of a simple pendulum of length $a$—call it the "swing" frequency—to the frequency $\sqrt{k/m}$ of an object of a mass $m$ bobbing on a spring of stiffness $k$—call it the "bounce" frequency. Using an available numeric processing program like IDL, MATLAB, OCTAVE, or PYTHON, explore the motions for several values of this one parameter, including values larger than, equal to, and smaller than 1. Write several paragraphs describing and presenting evidence for your discoveries.
Chapter 13

Evaluating Integrals

Frequently the answer to an interesting question in physics is given by—or can be cast in the form of—an integral, often as a function of the upper limit or as a parameter in the integrand. Sometimes, that integral can be evaluated in closed form. More often, however, the integral is analytically intractable and must be approached numerically. We begin this chapter by identifying several physical situations, the full addressing of which requires evaluating an integral, perhaps as a function of one or more parameters. Then we illustrate how to use symbolic algebra systems to approach those that can be evaluated analytically, describe a few of many available numerical algorithms (with attention to their accuracy), and—finally—describe ways to evaluate representative integrals using a variety of numerical approaches and computational tools. When parameters are involved, we also illustrate how to plot graphs of the integrals as functions of those parameters.

We shall classify each integral in one of three categories, since the approach to its numerical evaluation will depend on this classification. Integrals in the first category will simply be a number; their numerical evaluation involves a single invocation of one or another basic algorithm. In more complicated—and more interesting—cases, the integral will be a function of a parameter, which may appear in the limits of integration (second category) or embedded in the integrand (third category); the numerical evaluation of these integrals as a function of the parameter will involve repeated invocation of one or another basic algorithm within a loop.

13.1 Sample Problems

In this section, we identify several physical contexts in which integrals arise, and we determine a representative integral for each case.

13.1.1 One-Dimensional Trajectories

The motion of a particle of mass $m$ in one dimension under the action of a force $f$ satisfies Newton’s second law, equivalent to the two equations

$$\frac{dp}{dt} = f \quad \text{and} \quad \frac{dx}{dt} = v$$  \hspace{1cm} (13.1)

which are to be solved subject to the prescribed initial conditions $x(0) = x_0$, $v(0) = v_0$, and $p(0) = p_0$. Here, $x$, $v$, and $p$, are the position, velocity, and momentum of the particle, respectively, and, with $c$ standing for the speed of light, the relationship between $p$ and $v$ assumes one of the forms

$$p = \frac{mv}{\sqrt{1 - v^2/c^2}} \quad \text{or} \quad p = mv$$  \hspace{1cm} (13.2)
depending on whether the motion is relativistic or nonrelativistic. If the force happens to depend only on \( t \), the solution to these two differential equations can be expressed as the explicit integrals

\[
p(t) = p_0 + \int_0^t f(t') \, dt' \quad \text{and} \quad x(t) = x_0 + \int_0^t v(t') \, dt'
\]

(13.3)

and the \textit{physical} problem of predicting the trajectory reduces to the \textit{mathematical} problem of evaluating two integrals, finding the momentum from the first integral in Eq. (13.3), then solving for the velocity \( v(t) \) using the appropriate member of Eq. (13.2), and finally finding \( x(t) \) from the second integral in Eq. (13.3). Each integral is a function of at least one parameter—the upper limit \( t \)—and will also depend on additional quantities (e.g., \( m, c, \ldots \)) in the integrand (unless a dimensionless casting happens to suppress them). These integrals fall into either the second or the third of our three categories.

If, on the other hand, the force happens to depend only on \( x \), we can recast the computational task by noting first that

\[
\frac{d^2x}{dt^2} = \frac{dv}{dt} = \frac{dv}{dx} \frac{dx}{dt} = v \frac{dv}{dx}
\]

(13.4)

Then, Newton’s second law becomes

\[
m \frac{d^2x}{dt^2} = f(x) \implies mv \frac{dv}{dx} = f(x) \implies mv \, dv = f(x) \, dx
\]

(13.5)

Finally, by integrating this last expression from initial values to general values at some other time, we find that

\[
m \int_{v_0}^v v' \, dv' = \frac{1}{2} mv^2 - \frac{1}{2} mv_0^2 = \int_{x_0}^x f(x') \, dx'
\]

(13.6)

and the task of finding the velocity (as a function of \( x \)) is reduced to the straight-forward evaluation of an integral. Once that integral has been evaluated, we can then predict the position by exploiting the relationship

\[
\frac{dx}{dt} = v(x) \implies dt = \frac{dx}{v(x)} \implies \int_{t_0}^t dt' = t - t_0 = \int_{x_0}^x \frac{dx'}{v(x')}
\]

(13.7)

and we have reduced the second step in the solution to the evaluation of an integral as well. Both integrals deduced in this paragraph may have internal parameters, so they might fall into either the second or the third of our three categories.

13.1.2 Center of Mass

The location \( \mathbf{r}_{\text{cm}} \) of the \textit{center of mass} of an object is given by

\[
\mathbf{r}_{\text{cm}} = \frac{1}{M} \int \mathbf{r} \, dm
\]

(13.8)

where \( \mathbf{r} \) locates a representative element of the object, \( dm \) is the mass of that element, \( M = \int dm \) is the total mass of the object, and the integral extends over the region of space (volume, surface, or line) occupied by the object. More specifically, if the object lies in a plane and polar coordinates \((r, \phi)\) are appropriate, we would write this integral more explicitly in the form

\[
\mathbf{r}_{\text{cm}} = \frac{1}{M} \int \sigma(r, \phi) r \, r \, dr \, d\phi = \frac{1}{M} \int \sigma(r, \phi) \left( r \cos \phi \, \mathbf{i} + r \sin \phi \, \mathbf{j} \right) r \, dr \, d\phi
\]

(13.9)

where \( \sigma(r, \phi) \) is the mass per unit area of the object and \( r \, dr \, d\phi \) is the area of the chosen element.
Even more specifically, if we seek the center of mass of the uniform semicircular plate of total mass $M$, radius $a$, and mass per unit area $\sigma = M/(\frac{1}{2}\pi a^2)$ shown in Fig. 13.1, we would conclude that

$$r_{cm} = \frac{1}{M} \int_0^a \int_0^{\pi} \frac{M}{\frac{1}{2}\pi a^2} \left(r \cos \phi \hat{i} + r \sin \phi \hat{j}\right) r \, dr \, d\phi = \frac{2}{\pi a^2} \int_0^a \int_0^{\pi} r^2 \left(\cos \phi \hat{i} + \sin \phi \hat{j}\right) dr \, d\phi$$

(13.10)

Recognizing as always the wisdom of casting problems to be addressed with a computer in dimensionless form, we finally introduce the dimensionless length $\lambda = r/a$, in terms of which we then find that

$$\frac{r_{cm}}{a} = \frac{2}{\pi} \int_0^1 \int_0^{\pi} \lambda^2 \left(\cos \phi \hat{i} + \sin \phi \hat{j}\right) d\phi d\lambda$$

(13.11)

Although this integral is two-dimensional, it clearly falls into the first of our three categories—an integral whose value is simply a number (or, in this case, a constant vector).

### 13.1.3 Moment of Inertia; Radius of Gyration

The *moment of inertia* $I$ of an object of mass $M$ about a chosen axis is given by

$$I = \int r^2 \, dm$$

(13.12)

where $r$ is the distance of an element of the object from the chosen axis, $dm$ is the mass of that element, and the integral extends over the region of space (volume, surface, or line) occupied by the object. Further, the *radius of gyration* $k$ of this object with respect to the *same* axis is defined so that a point object of mass $M = \int dm$ located at the distance $k$ from the axis has the same moment of inertia as the object itself, i.e., $k$ is defined so that

$$I = Mk^2 \quad \implies \quad k = \sqrt{\frac{I}{M}}$$

(13.13)

Suppose, for example, we seek the moment of inertia of the uniform semicircular plate of mass $M$ and radius $a$ shown in Fig. 13.1 about the $x$ axis. As in the previous example, the mass per unit area $\sigma$ is given by $\sigma = M/(\frac{1}{2}\pi a^2)$. We choose a horizontal strip, all elements of which are the same distance $y$ from the $x$ axis. If this strip has mass $dm$, its contribution to the moment of inertia about that axis is $y^2 \, dm$, and the moment of inertia of the plate about that axis is given by

$$I = \int_{y=0}^{y=a} y^2 \, dm$$

(13.14)
CHAPTER 13. EVALUATING INTEGRALS

Figure 13.2: A simple pendulum.

Remembering that the equation of a circle of radius \(a\) is \(x^2 + y^2 = a^2\), we note next that the length of the illustrated strip at height \(y\) is \(2\sqrt{a^2 - y^2}\). If we take the width of that strip to be \(dy\), then its area is given by \(dA = 2\sqrt{a^2 - y^2} \, dy\), and its mass is given by

\[
dm = \sigma \, dA = \frac{4M}{\pi a^2} \sqrt{a^2 - y^2} \, dy
\] (13.15)

Finally, the moment of inertia of the entire object is given by the integral

\[
I = \frac{4M}{\pi a^2} \int_0^a y^2 \sqrt{a^2 - y^2} \, dy = \frac{4}{\pi} Ma^2 \int_0^1 \lambda^2 \sqrt{1 - \lambda^2} \, d\lambda
\]

\(\Rightarrow I = \frac{4}{\pi} \int_0^1 \lambda^2 \sqrt{1 - \lambda^2} \, d\lambda
\) (13.16)

where we have introduced the dimensionless variable \(\lambda = y/a\) and expressed the moment of inertia in units of \(Ma^2\), which is the moment of inertia of a point mass \(M\) a distance \(a\) from the axis. This integral is simply a number, and it therefore falls into the first of our three categories.

13.1.4 The Large Amplitude Simple Pendulum

Suppose we seek the period \(T\) of a simple pendulum of length \(l\) and mass \(m\) as shown in Fig. 13.2, but we do not wish to make the conventional small amplitude approximation, under which the period \(T_0\) is given by \(2\pi\sqrt{l/g}\), where \(g\) is the acceleration of gravity. With \(y\) standing for the vertical coordinate of the pendulum (measured upward from its point of support), \(I = ml^2\) for its moment of inertia about the point of support, and \(\omega\) for its angular velocity, we start by noting that the total energy (kinetic plus potential) of the pendulum in a general position is given by

\[
\text{energy} = \frac{1}{2}I\omega^2 + mgy = \frac{1}{2}ml^2 \left( \frac{d\theta}{dt} \right)^2 - mgl \cos \theta
\] (13.17)

If, with \(\theta_0\) symbolizing the amplitude of the motion, we invoke conservation of energy, equating the energy at a general point to the energy at the highest point in the swing (where \(d\theta/dt = 0\) and \(\theta = \theta_0\)), we find that

\[
\frac{1}{2}ml^2 \left( \frac{d\theta}{dt} \right)^2 - mgl \cos \theta = -mgl \cos \theta_0 \quad \Rightarrow \quad \frac{d\theta}{dt} = \pm \sqrt{\frac{2g}{l}} \sqrt{\cos \theta - \cos \theta_0}
\] (13.18)

(We assume that the pendulum does not have sufficient energy to swing over the top.) This last relationship then leads to the conclusion that

\[
dt = \pm \sqrt{\frac{l}{2g}} \frac{d\theta}{\sqrt{\cos \theta - \cos \theta_0}} \quad \Rightarrow \quad \int_0^{T/4} dt = \frac{T}{4} = \sqrt{\frac{l}{2g}} \int_{\theta_0}^0 \frac{d\theta}{\sqrt{\cos \theta - \cos \theta_0}}
\] (13.19)
where we have integrated over one-quarter of the period \((0 < t < T/4)\) in time and over one-quarter of the swing \((0 < \theta < \theta_0)\) in angle, and we have taken the positive sign because, in its motion over this interval, the pendulum indeed has positive angular velocity. Further, we have assumed from symmetry that the full period \(T\) is four times the time required for the pendulum to swing from its lowest point to its highest point.

This integral can be recast in numerous ways. Anticipating an existing standard function, we invoke the half angle identity \(\cos \theta = 1 - 2 \sin^2(\theta/2)\), finding that

\[
T = 2 \sqrt{\frac{l}{g}} \int_0^{\theta_0} \frac{d\theta}{\sqrt{\sin^2(\theta_0/2) - \sin^2(\theta/2)}}
\] (13.20)

We then introduce the variable \(\phi\) defined by \(\sin(\theta/2) = \sin(\theta_0/2) \sin \phi\) to find that

\[
T(k) = 4 \sqrt{\frac{l}{g}} \int_0^{\pi/2} \frac{d\phi}{\sqrt{1 - k^2 \sin^2 \phi}}
\] (13.21)

where the (dimensionless) parameter \(k = \sin(\theta_0/2)\) is determined by the amplitude of the pendulum’s motion. By writing \(T(k)\), we have drawn attention in the notation to the dependence of the period on \(k\) (and hence on the amplitude). Note also that, when the amplitude is small, \(k \approx 0\) and the integral can be readily evaluated to yield that \(T_0 = 2\pi \sqrt{l/g}\), which—reassuringly—agrees with the known value quoted at the beginning of this subsection. Then, expressing the period \(T(k)\) in units of \(T_0\), we find finally that

\[
\frac{T(k)}{T_0} = \frac{2}{\pi} \int_0^{\pi/2} \frac{d\phi}{\sqrt{1 - k^2 \sin^2 \phi}} = \frac{2}{\pi} K(k)
\] (13.22)

Here, for purposes of notation alone, we have introduced the integral

\[
K(k) = \int_0^{\pi/2} \frac{d\phi}{\sqrt{1 - k^2 \sin^2 \phi}}
\] (13.23)

which defines a standard, tabulated function known as the complete elliptic integral of the first kind. Note that the value of this integral depends on a parameter not in a limit but in the very structure of the integrand. This integral falls into the third of our three categories.

### 13.1.5 Statistical Data Analysis

When repeated measurements of a single quantity are subject to a large number of individually small, random fluctuations, the distribution of those measurements about their mean follows the normal or Gaussian distribution function given by

\[
G(x) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-(x-\mu)^2/(2\sigma^2)}
\] (13.24)

where \(\mu\) and \(\sigma\) are the mean and standard deviation of the distribution, respectively, and \(G(x)\) is normalized so that \(\int_{-\infty}^{\infty} G(x) \, dx = 1\). The probability that a single measurement will lie between \(\mu - \alpha\sigma\) and \(\mu + \alpha\sigma\) is then given by

\[
P(\mu - \alpha\sigma < x < \mu + \alpha\sigma) = \int_{\mu-\alpha\sigma}^{\mu+\alpha\sigma} G(x) \, dx = \frac{2}{\sqrt{\pi}} \int_0^{\alpha/\sqrt{2}} e^{-s^2} \, ds = \text{erf} \left( \frac{\alpha}{\sqrt{2}} \right)
\] (13.25)

where \(s = (x - \mu)/(\sigma\sqrt{2})\) and, for purposes of notation alone, we have introduced the integral

\[
\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-s^2} \, ds
\] (13.26)
which defines a standard, tabulated function known as the \textit{error function}. Thus, determining the probability given by Eq. (13.25) boils down to evaluating an integral that depends on a parameter appearing in the upper limit, i.e., to evaluating an integral that falls into the second of our three categories.

### 13.1.6 The Cornu Spiral

The two integrals

\[
C(u) = \int_0^u \cos \left( \frac{\pi t^2}{2} \right) \, dt \quad ; \quad S(u) = \int_0^u \sin \left( \frac{\pi t^2}{2} \right) \, dt \quad (13.27)
\]

which are called the \textit{Fresnel integrals}, appear in the study of Fresnel diffraction. Together, they define the \textit{Cornu spiral}, which is a graph of \(S(u)\) versus \(C(u)\). Each is a function of its upper limit as a parameter and falls into the second of our three categories.

### 13.1.7 Electric and Magnetic Fields and Potentials

Among the richest sources of important—and frequently analytically intractable—integrals are the relationships in electromagnetic theory that determine fields and potentials from prescribed sources. For a distribution of static charge, for example, we identify an element of that charge \(dq\) so that last find that

\[
(r,\varphi) = \text{a total charge and their centers separated by } 2a. \text{ The envisioned situation is shown in Fig. 13.3. Each ring carries a total charge } Q \text{ and, in mks units, we find the electric field } E(\mathbf{r}) \text{ and, in mks units, we find the electric field } E(\mathbf{r}) \text{ and the magnetic vector potential } A(\mathbf{r}) \text{ at the point } \mathbf{r} \text{ by evaluating the integrals}
\]

\[
E(\mathbf{r}) = \frac{1}{4\pi \epsilon_0} \int \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3} \, dq' \quad \text{and} \quad V(\mathbf{r}) = \frac{1}{4\pi \epsilon_0} \int \frac{dq'}{|\mathbf{r} - \mathbf{r}'|} \quad (13.28)
\]

which extend over all charges in the source. Similarly, for a source consisting of a steady current \(I'\) in a wire, we identify an element \(d\mathbf{r}'\) of the wire located at \(\mathbf{r}'\) and, in mks units, find the magnetic field \(B(\mathbf{r})\) and the magnetic vector potential \(A(\mathbf{r})\) at the point \(\mathbf{r}\) by evaluating the integrals

\[
B(\mathbf{r}) = \frac{\mu_0}{4\pi} \int \frac{I' \, d\mathbf{r}' \times (\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} \quad \text{and} \quad A(\mathbf{r}) = \frac{\mu_0}{4\pi} \int \frac{I' \, d\mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|} \quad (13.29)
\]

which extend over the path followed by the current.

More specifically, suppose we seek the electrostatic potential in the plane midway between two identical uniformly charged circular rings of radius \(a\) with their planes parallel, their axes coincident, and their centers separated by \(2a\). The envisioned situation is shown in Fig. 13.3. Each ring carries a total charge \(Q\) with (linear) charge density \(\lambda\). Using cylindrical coordinates \((r, \phi, z)\), we first find the potential at the general point \(\mathbf{r} = r \cos \phi \, \hat{i} + r \sin \phi \, \hat{j} + z \, \hat{k}\), produced by one such ring positioned in the \(xy\) plane with its center at the origin. Let \(\mathbf{r}' = a \cos \phi' \, \hat{i} + a \sin \phi' \, \hat{j}\) locate an element on that (single) ring. In this notation,

\[
\mathbf{r} - \mathbf{r}' = (a \cos \phi - a \cos \phi') \, \hat{i} + (a \sin \phi - a \sin \phi') \, \hat{j} + z \, \hat{k}
\]

(13.30) so

\[
|r - r'| = \left[ (r \cos \phi - a \cos \phi')^2 + (r \sin \phi - a \sin \phi')^2 + z^2 \right]^{1/2}
\]

(13.31)

Placing these results into the second member of Eq. (13.28) and recognizing that \(dq' = \lambda a \, d\phi'\), we at last find that

\[
V_{\text{one}}(r, \phi, z) = \frac{1}{4\pi \epsilon_0} \int_0^{2\pi} \frac{\lambda a \, d\phi'}{|r^2 + a^2 - 2ar \cos(\phi' - \phi) + z^2|^{1/2}} \quad (13.32)
\]
The entire integral assumes a simpler appearance, however, on the variable \( \alpha = \phi' - \phi \), becoming

\[
V_{\text{one}}(r, \phi, z) = \frac{\lambda a}{4\pi \varepsilon_0} \int_0^{2\pi} \frac{d\alpha}{\left[ r^2 + a^2 - 2ar \cos \alpha + z^2 \right]^{1/2}}
\]  

(13.33)

where we have invoked the periodicity of the integrand in \( \alpha \) to justify writing the integral to run from 0 to 2\( \pi \) rather than from \(-\phi\) to \(2\pi - \phi\). (The integral extends over an entire period of the integrand in either case.) As implied by the symmetry, the potential has turned out not to depend on \( \phi \).

We can now return to our original question, which asked about the potential in the midplane when two rings of the sort to which Eq. (13.33) applies have their centers separated by 2\( a \). The observation point in the midplane is thus a distance \( a \) “above” one of the rings and the same distance \( a \) “below” the other. We find the potential produced by these two rings by adding a contribution from each ring, concluding that

\[
V_{\text{midplane}}(r) = V_{\text{one}}(r, \phi, -a) + V_{\text{one}}(r, \phi, a) = \frac{2\lambda a}{4\pi \varepsilon_0} \int_0^{2\pi} \frac{d\alpha}{\left[ r^2 + 2a^2 - 2ar \cos \alpha \right]^{1/2}}
\]  

(13.34)

Finally, introducing the variable \( s \) defined by \( r = sa \) to express the radial coordinate in dimensionless terms, we find the expression

\[
V_{\text{two}}(s) = V_{\text{midplane}}(sa) = \frac{2\lambda a}{4\pi \varepsilon_0 a} \int_0^{2\pi} \frac{d\alpha}{\left[ s^2 - 2s \cos \alpha + 2 \right]^{1/2}}
\]  

(13.35)

or, even better, the expression

\[
V(s) = \frac{\sqrt{2}}{2\pi} \int_0^{2\pi} \frac{d\alpha}{\left[ s^2 - 2s \cos \alpha + 2 \right]^{1/2}}
\]  

(13.36)

where \( V(s) = V_{\text{two}}(s)/V_{\text{two}}(0) \). Equation (13.36) gives the potential at the radial coordinate \( r = sa \) in the midplane between two uniformly charged rings of radius \( a \) separated by 2\( a \). This integral falls into the third of our three categories, since its value is a function of the parameter \( s \) in the integrand.
13.1.8 Quantum Probabilities

In the standard interpretation, the wave function $\psi(x)$ for a one-dimensional quantum system is a probability amplitude, and the quantity $|\psi(x)|^2$ is a probability density. Further the wave function by convention is normalized so that $\int_{-\infty}^{\infty} |\psi(x)|^2 \, dx = 1$. Thus, the integral

$$P(x_1 < x < x_2) = \int_{x_1}^{x_2} |\psi(x)|^2 \, dx \quad (13.37)$$

gives the probability of finding the quantum system with its coordinate somewhere between $x = x_1$ and $x = x_2$.

More specifically, we remember that the wave function for a quantum harmonic oscillator in its ground state is given by

$$\psi(x) = \left(\frac{m\omega}{\pi\hbar}\right)^{1/4} e^{-y^2/2} \quad (13.38)$$

where $\hbar$ is Planck’s constant, $m$ is the mass of the oscillator and, with $k$ the spring constant and $a = \sqrt{\hbar\omega/k}$, $\omega = \sqrt{k/m}$ is the frequency of the oscillator; $y = x/a$ is a dimensionless coordinate.

Since the classical turning point of the oscillator occurs when its energy ($\frac{1}{2}\hbar\omega$ for the ground state) is equal to the potential energy $\frac{1}{2}kx^2$ turn, the turning point of this oscillator is given by

$$\frac{1}{2}\hbar\omega = \frac{1}{2}kx^2_{\text{turn}} \implies x_{\text{turn}} = \sqrt{\frac{\hbar\omega}{k}} = a \quad (13.39)$$

which provides a classical interpretation for the parameter $a$. The probability that the particle in the ground state of a quantum harmonic oscillator will be found in the classically forbidden region (i.e., somewhere outside the classical turning point) is given by the integral

$$P(|x| > |x_{\text{turn}}|) = \int_{-\infty}^{-a} |\psi(x)|^2 \, dx + \int_{a}^{\infty} |\psi(x)|^2 \, dx = 1 - \int_{-a}^{a} |\psi(x)|^2 \, dx \quad (13.40)$$

Finally, after substituting the wave function and recasting the entire integral as an integral on the variable $y$, we find that

$$P(|x| > |x_{\text{turn}}|) = 1 - \frac{1}{\sqrt{\pi}} \int_{-1}^{1} e^{-y^2} \, dy = 1 - \frac{2}{\sqrt{\pi}} \int_{0}^{1} e^{-y^2} \, dy = 1 - \text{erf}(1) \quad (13.41)$$

which contains no parameters and thus falls into the first of our three categories.

13.1.9 Expansion in Orthogonal Functions

Suppose we have identified a set of functions $\phi_i(x)$, $i = 1, 2, 3, \ldots$, that have the property

$$\int_{a}^{b} \phi_i(x) \phi_j(x) \, w(x) \, dx = N_j \delta_{ij} \quad (13.42)$$

where $N_j$ is a constant, $\delta_{ij}$ is the Kronecker delta (which has the value 1 when the indices are equal and the value 0 otherwise), $w(x)$ is a known weight function, and $a$ and $b$ define a known interval. These functions are said to be orthogonal with weight $w(x)$ on the interval $a \leq x \leq b$. The members of this set provide a basis in terms of which any arbitrary function $f(x)$ defined on the same interval can be expanded in the form

$$f(x) = \sum_{n} a_n \phi_n(x) \quad (13.43)$$
Though the argument we will here present is not mathematically rigorous, the expansion coefficients \( a_n \) can be quickly determined by multiplying Eq. (13.43) by \( w(x) \phi_j(x) \), integrating over the interval \( a \leq x \leq b \), and exchanging the order of integration and summation to find that

\[
\int_a^b w(x) \phi_j(x) f(x) \, dx = \sum_n a_n \int_a^b w(x) \phi_j(x) \phi_n(x) \, dx = \sum_n a_n N_n \delta_{nj} = N_j a_j \quad (13.44)
\]

We conclude that, if we know either of \( f(x) \) or \( a_n \), we can determine the other, i.e., that

\[
f(x) = \sum_n a_n \phi_n(x) \quad \iff \quad a_n = \frac{1}{N_n} \int_a^b w(x) \phi_n(x) f(x) \, dx \quad (13.45)
\]

The determination of the coefficients in this expansion of a known function clearly involves the evaluation of integrals, which explains the inclusion of this example in this chapter.

While many sets of orthogonal functions could be enumerated (see exercises), probably the most common set is\(^1\)

\[
\{1, \sin \frac{n\pi x}{l}, \cos \frac{m\pi x}{l}; \quad n, m = 1, 2, 3, \ldots \} \quad (13.46)
\]

Direct evaluation of the integral of each member of this set with all other members will reveal that these functions are orthogonal on the interval \( -l \leq x \leq l \) with weight \( w(x) = 1 \), i.e., that

\[
\int_{-l}^l 1 \times 1 \, dx = 2l \quad ; \quad \int_{-l}^l 1 \times \sin \frac{n\pi x}{l} \, dx = 0 \quad ; \quad \int_{-l}^l 1 \times \cos \frac{n\pi x}{l} \, dx = 0 \quad (13.47)
\]

\[
\int_{-l}^l \sin \frac{n\pi x}{l} \sin \frac{m\pi x}{l} \, dx = l \delta_{nm} \quad ; \quad \int_{-l}^l \sin \frac{n\pi x}{l} \cos \frac{m\pi x}{l} \, dx = 0 \quad (13.48)
\]

\[
\int_{-l}^l \cos \frac{n\pi x}{l} \cos \frac{m\pi x}{l} \, dx = l \delta_{nm} \quad ; \quad \int_{-l}^l \cos \frac{n\pi x}{l} \sin \frac{m\pi x}{l} \, dx = 0 \quad (13.49)
\]

The expansion\(^2\)

\[
f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left( a_n \cos \frac{n\pi x}{l} + b_n \sin \frac{n\pi x}{l} \right) \quad (13.50)
\]

of a function \( f(x) \) that is periodic with period \( 2l \) in this set of orthogonal functions is called a Fourier series. The coefficients are given by the integrals

\[
a_n = \frac{1}{l} \int_{-l}^l f(x) \cos \frac{n\pi x}{l} \, dx \quad ; \quad b_n = \frac{1}{l} \int_{-l}^l f(x) \sin \frac{n\pi x}{l} \, dx \quad (13.51)
\]

### 13.4 Evaluating Integrals Symbolically with Mathematica

The Mathematica command for symbolic integration is `Integrate`. The first argument in this command either is or stands for the integrand, and the second argument specifies the integration variable (with an optional range to convey limits). Thus, for example, the statement

\[
\text{Integrate}[\text{Sin}[k \times x], x]
\]

evaluates the indefinite integral \( \int \sin(kx) \, dx \) while the statement

\[
\text{Integrate}[\text{Sin}[k \times x], \{x, a, b\}]
\]

evaluates the definite integral \( \int_a^b \sin(kx) \, dx \) with limits \( a \) and \( b \).
Integrate[ Sin[k*x], {x, 0, Pi} ]

evaluates the \textit{definite} integral \( \int_0^\pi \sin(kx)\, dx \). If \textit{Mathematica} is unable to evaluate a particular integral, the program returns the noun form of the integral, i.e., the program simply returns the integral that was provided as input. \textit{Mathematica} also evaluates improper integrals, recognizing the symbols \textit{Infinity} for \(+\infty\) and \textit{-Infinity} for \(-\infty\). Whether integrating by hand or with the help of a symbolic program, we must always be wary of integrals when the integrand has a singularity in the interval of integration. Indeed, \textit{Mathematica} checks for these singularities automatically and prints a warning if evidence of a singularity is found.

In this section, we illustrate the use of the command \texttt{Integrate} to evaluate some of the integrals deduced in Section 13.1. Beyond integration \textit{per se}, we also show how, in many cases, \textit{Mathematica} can be used to set up the integral as well. To abbreviate the presentation of \textit{Mathematica} dialogs, we make liberal use of terminating semicolons to suppress intermediate output. You are therefore urged to duplicate the dialogs in an actual session with \textit{Mathematica}, removing semicolons so that intermediate output is explicitly displayed.

13.4.1 Relativistic Motion Under a Constant Force

Suppose the particle to which Eq. (13.3) applies moves relativistically, starting from rest at the origin \( [x(0) = 0, v(0) = 0, p(0) = 0] \), and experiences a \textit{constant} force \( f \). Then, the basic relationships from which we would determine \( x(t) \) and \( v(t) \) are

\[
\begin{align*}
p(t) &= \int_0^t f\, dt', \quad x(t) = \int_0^t v(t')\, dt', \quad \text{and} \quad p(t) = \frac{mv(t)}{\sqrt{1 - v(t)^2/c^2}} \quad (13.52)
\end{align*}
\]

The following “conversation” with \textit{Mathematica} will find the desired quantities and explore a few of their properties:

\begin{verbatim}
In[1]:= p = Integrate[f, {tp, 0, t}] Integrate (constant) f to find p.
Out[1]= ft

In[2]:= p == m * v / Sqrt[ 1 - v^2/c^2 ] Relate momentum to velocity.
Out[2]= ft == mv/\sqrt{1-v^2/c^2}

In[3]:= soln = Solve[ %, v ] Extract solution as an expression. (Physically, v and f must have the same sign.)
Out[3]= {{v -> \(-cft\) \sqrt{c^2m^2 + f^2t^2}}, {v -> \(cft\) \sqrt{c^2m^2 + f^2t^2}}}

In[4]:= v = soln[[2,1,2]]
Out[4]= \(\frac{cft}{\sqrt{c^2m^2 + f^2t^2}}\)

In[5]:= Limit[ v, t -> Infinity ]
Out[5]= \(\frac{cf}{\sqrt{f^2}}\)

In[6]:= Simplify[ %, f > 0 ]
Out[6]= c
\end{verbatim}

\footnote{\textit{Mathematica} statements are shown on the left; comments describing the statements are shown on the right.}
In[7]:= vp = v /. t -> tp;  
Substitute tp for t in v.
Integrate vp to find position.
In[8]:= x = Integrate[ vp, { tp, 0, t} , Assumptions->{c>0,m>0,t>0,f>0} ]
Out[8]=
\[\frac{c}{f}(-cm + \sqrt{c^2m^2 + f^2t^2})\]
Find limit of \(v\) for small \(t\).
In[9]:= Series[ v, { t, 0, 4 } ];
Out[10]=
\(v_{\text{classical}}\) = Simplify[ %, {c>0, m>0, f>0} ]
Out[11]=
\(\frac{ft}{m} - \frac{f^3t^3}{2c^2m^3} + O[t^5]\)
Find limit of \(x\) for small \(t\).
In[12]:= Series[ x, { t, 0, 5 } ];
In[13]=
\(x_{\text{classical}}\) = Simplify[ %, {c>0, m>0, f>0 } ]
Out[13]=
\(\frac{ft^2}{2m} - \frac{f^3t^4}{8c^2m^3} + O[t^6]\)
Obtain floating evaluation.

Reassuringly, the velocity approaches \(c\) as \(t \to \infty\) and, with the acceleration \(a\) identified as \(f/m\), \(x\) and \(v\) for small \(t\) (classical limit) are \(\frac{1}{2}at^2\) and \(at\) respectively.

13.4.2 Center of Mass

We next find the center of mass of the uniform, semicircular plate described in Section 13.1.2 as given by Eq. (13.9). We evaluate this integral with the statements

\(r_{\text{cm}}\) = \frac{scndint}{m}
Multiply each component by \(1/m\).

Since only the \(y\) component of \(r_{\text{cm}}\) differs from zero, the center of mass lies on the \(y\) axis—certainly not a surprise—a fraction \(4/3\pi = 0.4244\) of the radius from the center of the semicircle. Note, incidentally, that Mathematica has at lines In[4] and In[5] effected an element-by-element integration of the components of the list with a single statement.

Note also that we have evaluated the double integral as a sequence of single integrals. In the present case, the two integrals are independent of one another, and the order in which we perform the integrals is irrelevant. In some cases, however, the order may be important—as, for example, when the limits on one integration variable happen to depend on the other integration variable.
13.4.3 Moment of Inertia; Radius of Gyration

The evaluation by Mathematica of the moment of inertia of a semicircular plate as given by Eq. (13.16) is quick and straightforward. We need only the statements4

\[ I = \frac{m a^2}{4} \]

\[ k = \frac{a}{2} \]

to evaluate the necessary integral and find both the moment of inertia \( I = ma^2/4 \) and the radius of gyration \( k = a/2 \).

13.4.4 Electrostatic Potential of a Finite Line Charge

For a fourth example, we evaluate the electrostatic potential of a uniformly charged finite line extending along the \( z \) axis over the interval \(-a \leq z \leq +a\). The geometry is shown in Fig. 13.4. With \( \lambda \) representing the linear charge density on the line, \( dz' \) giving the length of an element of the line, \( r \) and \( r' \) locating the observation point and an element of the source, respectively, and \( dq' = \lambda dz' \), we deduce from Eq. (13.28) that the potential established by this source is given by

\[ V(x, y, z) = \frac{\lambda}{4\pi \epsilon_0} \int_{-a}^{+a} \frac{dz'}{|r - r'|} \]  

(13.53)

E lecting cylindrical coordinates \((r, \phi, z)\), we set \( r' = z' \hat{k} \) and \( r = r \cos \phi \hat{i} + r \sin \phi \hat{j} + z \hat{k} \). Then we construct the integrand and evaluate the integral with the statements5

\[ V(r, z) = \frac{\lambda}{4\pi \epsilon_0} \sqrt{r^2 + (z - z')^2} \]

Then, we find \( V(r, z) \), with \( r = \sqrt{x^2 + y^2} \) by executing the statement

---

4 The variable \( I \) in Mathematica is reserved for \( \sqrt{-1} \), so we use \( II \) for the moment of inertia.

5 The file \$HEAD/mathematica/crossdot.m was described in Section 8.11 and exists in the public directory structure. (The symbol \$HEAD is defined in the Local Guide.) It defines the two functions \( \text{lucross}[v1, v2] \) and \( \text{ludot}[v1, v2] \) for evaluating the cross and dot products of two three-component vectors when those vectors are conveyed by lists.
Figure 13.4: A line charge lying on the $z$ axis.

\[
\begin{align*}
\text{In[10]} & := V = \text{Integrate[ integ, \{ zp, -a, a \}, Assumptions -> \{ a>0, z>0, r>0 \} ]} \\
\text{Out[10]} & = \frac{\lambda \log \left[ \frac{a + z + \sqrt{r^2 + (a + z)^2}}{-a + z + \sqrt{r^2 + (-a + z)^2}} \right]}{4\pi\varepsilon_0}
\end{align*}
\]

While correct, this result is certainly not particularly transparent. Let us test it by examining its behavior in the $xy$ plane in the two limits $a/r \ll 1$ and $a/r \gg 1$, i.e., when $r$ is large compared to the length of the line and when $r$ is small compared to the length of the line. For the first limit, we execute the statements

\[
\begin{align*}
\text{In[11]} & := V_0 = V /. z \rightarrow 0 \\
\text{Out[11]} & = \frac{\lambda s}{2\pi\varepsilon_0} + O[s^3] \\
\text{In[12]} & := V_0 /. a \rightarrow s r; \\
\text{In[13]} & := \text{Series[ \%, \{ s, 0, 2 \} ];} \\
\text{In[14]} & := \text{Simplify[ \%, r>0 ]} \\
\text{Out[14]} & = \frac{\lambda s}{2\pi\varepsilon_0} + O[s^3]
\end{align*}
\]

Let $a/r \rightarrow s$.

\[
\begin{align*}
\text{In[15]} & := \text{Normal[\%];} \\
\text{In[16]} & := \% /. s \rightarrow a/r; \\
\text{In[17]} & := \% /. [\lambda] \rightarrow q/(2a) \\
\text{Out[17]} & = \frac{q}{4\pi\varepsilon_0 r}
\end{align*}
\]

As expected, the potential at points remote from the charged line varies with distance like that of a point charge. From a point far away from the wire (compared to its length), the wire looks like a point charge.
At the other extreme, when $r$ is small compared to the length of the line, we might think to invoke a statement like \( \text{Series}[V_0, \{r,0,2\}] \). That approach, however, is problematic because the denominator under the logarithm in \( V_0 \) is zero at \( r = 0 \) when \( a > 0 \). The expression \( V_0 \) does not have a very well behaved series expansion about the point \( r = 0 \). As an alternative, we might notice that the numerator and denominator under the logarithm in \( V_0 \) are well behaved in the vicinity of \( r = 0 \). In particular

\[
\text{In[18]} = \text{Series}[-a + \sqrt{a^2 + r^2}, \{r,0,2\}, \text{Assumptions} \rightarrow a > 0];
\]

\[
\text{Out[18]} = \frac{r^2}{2a}
\]

\[
\text{In[19]} = \text{Series}[a + \sqrt{a^2 + r^2}, \{r,0,2\}, \text{Assumptions} \rightarrow a > 0];
\]

\[
\text{Out[19]} = 2a + \frac{r^2}{2a}
\]

Substituting these expansions into \( V_0 \) with the statements

\[
\text{In[20]} = \text{SmallR} = V_0 /\{ -a + \sqrt{a^2 + r^2} \rightarrow \text{T1}, a + \sqrt{a^2 + r^2} \rightarrow \text{T2} \}
\]

\[
\text{Out[20]} = \frac{\lambda}{4\pi\epsilon_0} \log \left( \frac{2a(2a + \frac{r^2}{2a})}{r^2} \right)
\]

we find a much simpler, though approximate expression. Further, we can now see why a well-behaved series expansion does not exist. As \( r \) becomes ever smaller, the denominator in this logarithm approaches zero and the logarithm diverges. The numerator, however, approaches \( 4a^2 \) as \( r \rightarrow 0 \). Thus, we can simplify the expression further by ignoring the term \( \frac{r^2}{2a} \) in the numerator. The statement

\[
\text{In[22]} = \text{SmallR} = \text{SmallR} /\{ 2a + \sqrt{2}/(2a) \rightarrow 2a \}
\]

\[
\text{Out[22]} = \frac{\lambda}{4\pi\epsilon_0} \log \left( \frac{4a^2}{r^2} \right)
\]

\[
\text{In[22]} = \text{Quit}[\]

effects this further simplification. Indeed, if we recognize the equivalent form,

\[
-\frac{\lambda}{2\pi\epsilon_0} \ln \frac{r}{a} + \frac{\lambda \ln 2}{2\pi\epsilon_0}
\]

we see that our result differs by a constant from the familiar logarithmic result for the potential of a uniformly charged, infinitely long straight wire when the reference point (point of zero potential) is taken at a distance \( a \) from the wire. From a point close to the wire (compared to its length) and near its center, the wire looks to be infinitely long.

### 13.4.5 The Helmholtz Coil

Next, we evaluate the magnetic field produced on the \( z \) axis by a pair of circular current loops of arbitrary separation and demonstrate the significance of the specific separation used in the Helmholtz coil. Strategically, we regard the pair as a superposition of two loops and seek first the magnetic field of a single loop. In general, the magnetic field \( B(r) \) is given by the Biot-Savart law, which is the first
member of Eq. (13.29). If the (single) loop has radius \( a \) and lies in the \( xy \) plane with its center at the origin, then, for an observation point on the \( z \) axis, the various vectors in the expression are given by \( \mathbf{r} = s a \hat{k} \), where (to facilitate expressing things in dimensionless form) the \( z \) coordinate is written as a multiple \( s \) of the radius \( a \) of the loop, \( \mathbf{r}' = a \cos \phi \hat{i} + a \sin \phi \hat{j} \), and \( d\mathbf{r}' = (-a \sin \phi \hat{i} + a \cos \phi \hat{j}) d\phi \).

Evaluation of the integrand and the integral for a single loop then proceeds with the statements

\begin{verbatim}
In[1]:= Get[ "$HEAD/mathematica/crossdot.m" ];
In[2]:= r = { 0, 0, s * a };
In[3]:= rp = {a*Cos[\[Phi]],a*Sin[\[Phi]],0};;
In[4]:= drp = {-a*Sin[\[Phi]],a*Cos[\[Phi]],0};;
In[5]:= sep = r - rp;
In[6]:= ludot[ sep, sep ];
In[7]:= mag2 = Simplify[ % ];
In[8]:= lucross[ drp, sep ];
In[9]:= num = Simplify[ % ];
In[10]:= integ = num*\[Mu]0*i/(4*Pi*mag2^(3/2));
In[11]:= integ = Simplify[ %, a>0 ];
In[12]:= B1 = Integrate[ integ, { \[Phi], 0, 2*Pi } ];
In[13]:= B1z = B1[[3]];
In[14]:= B1c = B1z /. s -> 0;
In[15]:= B1z = B1z / B1c
Out[15]= 1
Out[16]= 1/2(1 + c^2)^(3/2) \\
\end{verbatim}

At the beginning, we were interested not in the field of a single loop but in the field of a pair of loops. Thus, we next combine the fields of two separate loops, one positioned a distance \( ca \) above the midplane and the other positioned a distance \( ca \) below the midplane, and again normalize the field so that it is measured in units defined by its value at the origin midway between the two loops. These objectives are accomplished with the statements

\begin{verbatim}
In[16]:= Bpair = (B1z /. s -> s-c) + (B1z /. s -> s+c);
In[17]:= Bpairc = Bpair /. s -> 0;
In[18]:= Bpair = Bpair / Bpairc
Out[18]= 1/2(1 + c^2)^(3/2) \\
\end{verbatim}

\footnote{See footnote 5.}
Figure 13.5: On-axis magnetic field of a pair of loops. The highest of the three graphs shows the field when the loop separation is $c = 1.0$ (separation of the two loops equal to twice the radius of each loop). The middle graph corresponds to $c = 0.5$ (the Helmholtz separation) and the lower graph corresponds to $c = 0.25$. This graph was created by Mathematica.

This expression is complicated. We explore it further in two ways. First, let us plot a few graphs of this result as a function of $s$ for representative values of $c$ with the statements

\begin{verbatim}
In[19]:= c100 = Bpair /. c -> 1.0;
In[20]:= c050 = Bpair /. c -> 0.5;
In[21]:= c025 = Bpair /. c -> 0.25;
In[22]:= Plot[{ c100, c050, c025 }, {s, 0.0, 3.0},
        PlotRange -> {0.0, 2.0}, PlotStyle -> Thickness[0.005],
        AxesLabel -> {"s", "B/Bc"} ]
\end{verbatim}

The resulting graph is shown in Fig. 13.5. As will be further supported in the next paragraph, the middle of the three graphs, which corresponds to $c = 0.5$ (separation of coils equal to the radius of each coil), has the largest region of near constant field near the on-axis point midway between the current loops (i.e., near $s = 0$).

Second, let us examine the behavior of the field near the center ($s = 0$) more closely. To do so, we suppose $s \to 0$ and look at the Taylor expansion of the field about $s = 0$ using the statement

\begin{verbatim}
In[23]:= tay = Simplify[ Series[ Bpair, { s, 0, 4 } ] ]
\end{verbatim}

\begin{verbatim}
Out[23]= 1 + 3(-1 + 4c^2)s^2 + 15(1 - 12c^2 + 8c^4)s^4 + O[s]^5
\end{verbatim}

Clearly there is a particular value of $c$—half the loop separation—at which the coefficient of the $s^2$ term is zero! For that separation, the magnetic field near the center of the loops is especially constant. We find that special value of $c$ with the statements\(^7\)

\[^7\text{Remember the structure of the storage of Taylor series as described in Section 8.8.7.}\]
In[24]:= Numerator[ Normal[tay][[2]] ]
Out[24]= 3(-1 + 4c^2)s^2
In[25]:= Solve[ % == 0, c ]
Out[25]= \{ \{ c \to -\frac{1}{2} \}, \{ c \to \frac{1}{2} \} \}

The first of these statements extracts the numerator of the second term of the Taylor series and
the second finds the value of \( c \) that makes that numerator zero. The two solutions are, of course,
equivalent; both reveal that the physical arrangement in which the field has no \( s^2 \) term has one
loop positioned one-half of the loop radius above the midplane and the other positioned one-half of
the loop radius below the midplane. That is, the loops are separated by their common radius—the
Helmholtz separation.

When the loops have that special separation, we determine numerically how the field varies
with \( s \) by using the statements
In[26]:= tay = Normal[ tay ]
Out[26]= 1 + 3(-1 + 4c^2)s^2 + \frac{15(1 - 12c^2 + 8c^4)s^4}{2(1 + c^2)^2} + \frac{8(1 + c^2)^4}{8(1 + c^2)^4}
In[27]:= tay /. c -> 1/2
Out[27]= 1 - \frac{144}{125}s^4
In[28]:= tay1 = \% // N
Out[28]= 1. - 1.152s^4

and we see again that, so long as we don’t stray too far from the center of the arrangement, the
magnetic field falls away from its central value as the fourth power of the distance from the center.
In particular (see next paragraph for a refinement), we might determine approximately how far we
can stray from the center along the axis before the field has fallen to, say, 99% or 95% of its value
at the center by executing the statements
In[29]:= Solve[ tay1 == 0.99, s ]
Out[29]= \{ \ldots, \{ s \to 0.305237 \}, \ldots \}
In[30]:= Solve[ tay1==0.95, s ]
Out[30]= \{ \ldots, \{ s \to 0.456435 \}, \ldots \}

Unfortunately, these solutions are only approximate, since we have used a truncated expansion of
the field, an expansion that becomes increasingly incorrect as \( s \) increases.

To find more accurately how far from the center we can stray before the field has fallen to 99%
or 95% of its value at the center, we need return to the exact expression of \( B_{\text{pair}} \) but restrict it to
the Helmholtz case by setting \( c = 1/2 \) with the statement

In[31]:= Bhelm = Bpair /. c -> 1/2:

Then, we seek the values of \( s \) at which this expression assumes the value 0.99 (1% fall off) or 0.95 (5%
fell off). Equivalently, we seek the roots of the expressions \( Bhelm - 0.99 = 0 \) and \( Bhelm - 0.95 = 0 \).
The dependence of these expression on \( s \), however, is complicated, and the command \( \text{Solve} \) either
takes a very long time or fails altogether. Thus, we draw (prematurely—see Section 14.13) on
Mathematica’s command \( \text{FindRoot} \), which uses an iterative numerical procedure to refine a supplied
initial guess for the root. The above results provide us with sensible initial guesses, so we invoke the
statements\(^8\)

\(^8\)In \( \text{FindRoot} \), the first argument is the equation whose root is sought and the second is a list whose first element is
In[32]:= FindRoot[ Bhelm - 0.99 == 0, { s, 0.3 } ]
Out[32]= 
{ s → 0.313746 }  
Find s when field is down 1%.

In[33]:= FindRoot[ Bhelm - 0.95 == 0, { s, 0.5 } ]
Out[33]= 
{ s → 0.488455 }  
Find s when field is down 5%.

Finally, out of curiosity, we ask about the value of the magnetic field at a point in the center of one of the coils by executing the statement

In[34]:= N[ Bhelm /. s -> 1/2 ]
Out[34]= 0.945824

In[35]:= Quit[]

In summary, as s increases (i.e., as the observation point moves away from the center of the Helmholtz pair), the on-axis field falls away from its value at the center, at least initially, by an amount proportional to $s^4 = (z/a)^4$. As the above results reveal, we would have to move 31.4% of the radius (62.8% of the distance from the center to the plane of either loop) before the field has fallen to 99% of its value at the center, and 48.8% of the radius (97.6% of the distance from the center to the plane of either loop) before the field has fallen to 95% of its value at the center. By the time we reach the plane of either loop, the field has fallen only to 94.6% of its value at the center. This optimal configuration of a pair of current loops is regularly used to produce a nearly uniform field over a large volume of space.

13.4.6 Period of a Pendulum: Correction as Amplitude Grows

The integral in Eq. (13.22) does not have a simple or familiar evaluation in closed form. Mathematica does, however, make it easy to examine the way in which the period departs from its limiting value as the amplitude moves away from the very small. We simply determine a Taylor expansion of the integrand with the statements

In[1]:= integ = 1/Sqrt[ 1-k^2*Sin[\[Phi]]^2 ]
Out[1]= \[\frac{1}{\sqrt{1-k^2 \cdot \text{Sin}^2(\phi)}}\]

In[2]:= intapprox = Series[ integ, {k, 0, 5} ]
Out[2]= 1 + \frac{1}{2} \text{Sin}^2(\phi)k^2 + \frac{3}{8} \text{Sin}^4(\phi)k^4 + O[k]^6

Then we insert the pre-multiplying factor $2/\pi$ and integrate the series on $\phi$ with the statement

In[3]:= Expand[ 2*Integrate[ intapprox, { \[Phi], 0, Pi/2 } ] / Pi ]
Out[3]= 1 + \frac{1}{4} k^2 + \frac{9}{64} k^4

Finally, we substitute $\sin(\theta_0/2)$ for $k$ with the statement

the variable to be adjusted in finding the root and whose second element is an initial guess for the value of the root. With this syntax, FindRoot invokes Newton’s method for finding the root. By default, FindRoot strives to find the root so that the function at the root is zero to within $\pm 10^{-6}$—though this goal is sometimes machine and installation dependent.
13.4. EVALUATING INTEGRALS SYMBOLICALLY WITH MATHEMATICA

In[4]:= % /. k -> Sin[\[Theta]0/2]

and expand this result as a power series in \[Theta]0

In[5]:= Normal[ Series[ %, { \[Theta]0, 0, 5 } ] ]

Note that \[Theta]0 must here be expressed in radians.

To see how significantly the period departs from the small-amplitude value when \[Theta]0 is not quite zero, let us evaluate this expression for \[Theta]0 = 5\degree, 10\degree, 15\degree, and 20\degree. We use the Mathematica statements

In[6]:= deg = { 5, 10, 15, 20 };                Set angles in degrees.
In[7]:= ampl = deg * (Pi/180);                Convert angles to radians.
Out[8]= {1.00048, 1.00191, 1.0043, 1.00767}

Evidently, at amplitudes of 5\degree, 10\degree, 15\degree, and 20\degree, the period of the pendulum is approximately 0.05\%, .2\%, .4\%, and .8\% larger than the standard small-amplitude approximation given by the expression 2\pi√(l/g).

13.4.7 Fourier Coefficients for Half-Rectified Signal

Suppose we sought to express the half-rectified sine wave

\[ f(x) = \begin{cases} 
0 & -l \leq x < 0 \\
\sin \frac{\pi x}{l} & 0 \leq x < l 
\end{cases} \] (13.54)

in a Fourier series. In accordance with Eq. (13.51), the coefficients would then be given by the integrals

\[ a_0 = \frac{1}{l} \int_{-l}^{0} \sin \frac{\pi x}{l} \, dx \]
\[ a_n = \frac{1}{l} \int_{-l}^{0} \sin \frac{\pi x}{l} \cos \frac{n\pi x}{l} \, dx \]
\[ b_n = \frac{1}{l} \int_{0}^{l} \sin \frac{\pi x}{l} \sin \frac{n\pi x}{l} \, dx \] (13.55)

where \( n = 1, 2, 3, \ldots \). We find these coefficients quickly. For \( a_0 \), we submit the statement

In[1]:= a0 = Integrate[ Sin[Pi*x/l], { x, 0, 1 } ]/l
Out[1]= 2/\[Pi]

Then for \( a_n \) when \( n \) is an integer not equal to zero, we use the statements

In[2]:= an = Integrate[ Sin[Pi*x/l]*Cos[n*Pi*x/l], { x, 0, 1 } ]/l;
In[3]:= an = Simplify[ an, Element[ n, Integers ] ]
Out[3]= -\frac{1 + (-1)^n}{\[Pi] - n^2\[Pi]}

Note, however, that this expression for \( n = 1 \) is indeterminate, so we evaluate that case in isolation with the statement

\(^9^\)Note the distinction between 1 (el) and 1 (one). Here and in the next several statements, the character is el.
Finally, we find \( b_n \) with the statement

\[
\text{In}[5]:= \text{bn} = \text{Integrate}[ \text{Sin}[\text{Pi}*x/\text{l}]*\text{Sin}[n*\text{Pi}*x/\text{l}], \{x, 0, \text{l}\}]/\text{l};
\]

\[
\text{In}[6]:= \text{bn} = \text{Simplify}[\% , \text{Element}[n, \text{Integers}]]
\]

\[
\text{Out}[6]= 0
\]

A close look at the integral giving \( b_n \), however, reveals that this result cannot be correct for \( b_1 \), since the integrand in that case \( \sin^2(\pi x/\text{l}) \) is decidedly positive and an integrand that is everywhere positive cannot yield zero for the integral.\(^\text{10}\) The formal evaluation of the integral (see the output of line In[5] on your screen) is, in fact, indeterminate at \( n = \pm 1 \). Rather than trying to evaluate that indeterminacy as a limit, we simply redo the integral with the statement

\[
\text{In}[7]:= \text{b1} = \text{Integrate}[\text{Sin}[\text{Pi}*x/\text{l}]^2, \{x, 0, \text{l}\}]/\text{l}
\]

\[
\text{Out}[7]= \frac{1}{2}
\]

In total, we find that

\[
a_0 = \frac{2}{\pi} \quad ; \quad a_1 = 0 \quad ; \quad a_n = -\frac{1+(-1)^n}{\pi(n^2-1)}, n > 1 \quad ; \quad b_1 = \frac{1}{2} \quad ; \quad b_n = 0, n > 1
\]

(13.56)

With these values for the Fourier coefficients, we assemble the Fourier series in accordance with Eq. (13.50) to find that

\[
f(x) = \frac{1}{\pi} - \sum_{n=2}^{\infty} \frac{1+(-1)^n}{\pi(n^2-1)} \cos \frac{n\pi x}{\text{l}} + \frac{1}{2} \sin \frac{\pi x}{\text{l}}
\]

\[
= \frac{1}{\pi} + \frac{1}{2} \sin \frac{\pi x}{\text{l}} - \frac{2}{\pi} \sum_{m=1}^{\infty} \frac{1}{4m^2-1} \cos \frac{2m\pi x}{\text{l}}
\]

\[
= \frac{1}{\pi} + \frac{1}{2} \sin \frac{\pi x}{\text{l}} - \frac{2}{\pi} \left( \frac{1}{3} \cos \frac{2\pi x}{\text{l}} + \frac{1}{15} \cos \frac{4\pi x}{\text{l}} + \frac{1}{35} \cos \frac{6\pi x}{\text{l}} + \cdots \right)
\]

(13.57)

In the second and third forms, we have recognized that \( 1+(-1)^n \) is zero when \( n \) is odd and 2 when \( n \) is even, so the sum on \( n \) can be extended over only even values of \( n \). Thus, setting \( n = 2m \), we have arranged in the second form for \( m \) to range over all positive integers.

The nature of this series is more clearly evident in a succession of graphs displaying the truncated series

\[
f_n(x) = \frac{1}{\pi} + \frac{1}{2} \sin \frac{\pi x}{\text{l}} - \frac{2}{\pi} \sum_{m=1}^{n} \frac{1}{4m^2-1} \cos \frac{2m\pi x}{\text{l}}
\]

(13.58)

which approaches \( f(x) \) as \( n \to \infty \). Introducing the variable \( \overline{x} = x/\text{l} \) (and then dropping the overbar), we then generate graphs for various values of \( n \). First, we define a function from which the coefficients \( a_m \) can be evaluated for various \( m \) with the statement

\[
\text{In}[8]:= \text{a}[m_] := \text{an} /. n -> 2*m
\]

\(^{10}\)Note also that Mathematica gave us no warning that the result displayed was potentially incorrect for \( n = \pm 1 \).
Figure 13.6: Succession of truncated series representing half-rectified sine wave. The upper left, upper right, lower left, and lower right graphs show the series when truncated at \( n = 1, 2, 4, \) and 8, respectively. This graph was created by *Mathematica*.

Then, we submit the statement

\[
\text{In}[9]:= f[x_,n_] := \frac{1}{\pi} + \sin(\pi x) / 2 + \sum_{m=1}^{n} a[m] \cos(2m\pi x), \quad \{ m, 1, n \} 
\]

to define a function for evaluating the truncated series as a function of \( x \) for selected \( n \). Finally, we plot graphs of the truncated series for \( n = 1, 2, 4, \) and 8 with the statements

\[
\text{In}[10]:= \text{Plot}[ f[x,1], \{x,-3.0,3.0\}, \text{PlotRange}\to\{-0.5,1.5\}, \\
\text{PlotStyle}\to\text{Thickness}[0.005], \text{AxesLabel}\to\{"x","f(x,1)"\} ] 
\]

\[
\text{In}[11]:= \text{Plot}[ f[x,2], \{x,-3.0,3.0\}, \text{PlotRange}\to\{-0.5,1.5\}, \\
\text{PlotStyle}\to\text{Thickness}[0.005], \text{AxesLabel}\to\{"x","f(x,2)"\} ] 
\]

\[
\text{In}[12]:= \text{Plot}[ f[x,4], \{x,-3.0,3.0\}, \text{PlotRange}\to\{-0.5,1.5\}, \\
\text{PlotStyle}\to\text{Thickness}[0.005], \text{AxesLabel}\to\{"x","f(x,4)"\} ] 
\]

\[
\text{In}[13]:= \text{Plot}[ f[x,8], \{x,-3.0,3.0\}, \text{PlotRange}\to\{-0.5,1.5\}, \\
\text{PlotStyle}\to\text{Thickness}[0.005], \text{AxesLabel}\to\{"x","f(x,8)"\} ] 
\]

\[
\text{In}[14]:= \text{Quit}[] 
\]

Shown in Fig. 13.6, the resulting graphs clearly reveal the convergence of the sum to the half-rectified wave as more terms are included.
13.5 Algorithms for Numerical Integration

Unfortunately, very many integrals of great interest have no closed form, analytic evaluation. To address these integrals, numerical analysts have developed many formulae—often called quadrature formulae—for numerical integration. In this section, we describe several of these formulae.

13.5.1 Newton-Cotes Quadrature

One family of quadrature formulae can be deduced by starting with the recognition that the definite integral

\[ A = \int_{a}^{b} f(x) \, dx \]  

(13.59)

represents geometrically the area under the graph of \( f(x) \) over the interval \( a \leq x \leq b \). As shown in Fig. 13.7, let that interval be divided into \( N \) segments, each of width \( \Delta x = (b - a)/N \), let \( x_0 = a \), \( x_1 = a + \Delta x \), \( x_2 = a + 2 \Delta x \), \ldots, \( x_i = a + i \Delta x \), \ldots, \( x_N = b \), and let \( f(x_0) = f(a) = f_0 \), \( f(x_1) = f_1 \), \ldots, \( f(x_i) = f_i \), \ldots, \( f(x_N) = f(b) = f_N \). To deduce the simplest quadrature formula, we approximate the area of each resulting strip by the area of a rectangle whose height is the value of \( f(x) \) at the left end of the strip [Fig. 13.8(a)]. Thus

\[
\int_{a}^{b} f(x) \, dx \approx f_0 \Delta x + f_1 \Delta x + \cdots + f_i \Delta x + \cdots + f_{N-1} \Delta x
\]

(13.60)

which turns out to be 100% accurate only if \( f(x) \) happens to be a constant.

If, however, we approximate the area of each strip by the area of a rectangle whose height is the value of \( f(x) \) at the midpoint of the strip, we would deduce the midpoint rule

\[
\int_{a}^{b} f(x) \, dx \approx M_N = \left( f_{1/2} + f_{3/2} + f_{5/2} + \cdots + f_{N-3/2} + f_{N-1/2} \right) \Delta x
\]

(13.61)

which turns out to be 100% accurate when \( f(x) \) is a linear function of \( x \). (When \( f(x) \) is linear, the error made by overestimating the function in one half of the interval is exactly compensated by the error made by underestimating the function in the other half of the interval. The formula turns out to be 100% accurate for a polynomial of one higher order than the polynomial used—here a constant—to approximate the function in each strip.)

For a further refinement, we might approximate the area of each strip by the area of a trapezoid [Fig. 13.8(b)], in which case we obtain the trapezoidal rule,

\[
\int_{a}^{b} f(x) \, dx \approx T_N = \frac{1}{2} (f_0 + f_1) \Delta x + \frac{1}{2} (f_1 + f_2) \Delta x + \cdots + \frac{1}{2} (f_{N-1} + f_N) \Delta x
\]

\[
= \left( \frac{1}{2} f_0 + f_1 + f_2 + \cdots + f_i + \cdots + f_{N-1} + \frac{1}{2} f_N \right) \Delta x
\]
13.5. Algorithm for Numerical Integration

Figure 13.8: Two different approximations leading to quadrature formulae. The left figure corresponds to Eq. (13.60), the right figure to Eq. (13.62).

\[ \int_a^b f(x) \, dx \approx \frac{1}{3} \left( f(a) + 4f(x_1) + f(b) \right) \Delta x \]

which, as with the midpoint rule, is 100% accurate when \( f(x) \) is linear, i.e., when \( f(x) \) is a polynomial of the same order as the one used to approximate the function.

A final (for here) and still better approximation is obtained if we pair the strips—which then requires \( N \) to be even—and approximate the area of each pair by the area under the parabola fitted to the values of \( f(x) \) at the three points defining the pair. The result,

\[ \int_a^b f(x) \, dx \approx S_N = \frac{1}{3} (f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots + 2f_{N-2} + 4f_{N-1} + f_N) \Delta x \]

is called Simpson’s rule. (See the first exercise in Section 13.15.2.) For reasons similar to those that apply for the midpoint rule, Simpson’s rule is 100% accurate for cubic polynomials, one order higher than the quadratic polynomial used to approximate the function.

Continuation of this procedure to higher and higher degree polynomials generates a succession of increasingly more accurate—but also more and more complicated—Newton-Cotes formulae, which are characterized in particular by evaluating the function at equally spaced interpolation points.

13.5.2 Rearrangements for Computational Efficiency

Two rearrangements of the formulae deduced to this point facilitate the writing of more efficient algorithms. Suppose, for example, that we use the trapezoidal rule and write out a succession of formulae for evaluating \( \int_a^b f(x) \, dx \) for \( N = 1, 2, 4, 8, 16, \ldots \) divisions of the interval of integration. We find first that

\[ T_1 = T_{2^0} = \frac{f(a) + f(b)}{2} (b - a) \]

Then, halving the step size and introducing \( x_1 = a + (b - a)/2 \), the midpoint of the interval (see Fig. 13.9, in which—for present convenience—we label the points differently than we did in Fig. 13.7), we find that

\[ T_2 = T_{2^1} = \left[ \frac{f(a)}{2} + f(x_1) + \frac{f(b)}{2} \right] \frac{(b - a)}{2} = \frac{1}{2} T_1 + f(x_1) \frac{(b - a)}{2} \]

\[ = \frac{1}{2} T_1 + f \left( a + \frac{b - a}{2} \right) \frac{(b - a)}{2} \]

(13.65)
CHAPTER 13. EVALUATING INTEGRALS

Figure 13.9: A rearrangement of the labels for dividing points.

Halving the step size again and introducing \( x_2 = a + (b - a)/4 \) and \( x_3 = a + 3(b - a)/4 \) (again, see Fig. 13.9), we find that

\[
T_4 = T_2^2 = \left[ \frac{f(a)}{2} + f(x_2) + f(x_1) + f(x_3) + \frac{f(b)}{2} \right] \frac{(b - a)}{4}
\]

\[
= \frac{1}{2} T_2 + \left[ f(x_2) + f(x_3) \right] \frac{(b - a)}{4}
\]

\[
= \frac{1}{2} T_2 + \left[ f \left( a + \frac{b - a}{4} \right) + f \left( a + 3 \frac{b - a}{4} \right) \right] \frac{(b - a)}{4}
\]

(13.66)

Yet again, introducing

\[
x_4 = a + \frac{1}{8} (b - a) \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \q
and that
\[ T_8 = \left[ \frac{1}{2} f(a) + f(x_4) + f(x_2) + f(x_5) \right. \\
+ f(x_1) + f(x_6) + f(x_3) + f(x_7) + \left. \frac{1}{2} f(b) \right] \frac{b-a}{8} \] (13.72)

Note, in particular, the combination
\[ 4T_8 - T_4 = \left[ f(a) + 4f(x_4) + 2f(x_2) + 4f(x_5) + 2f(x_1) \right. \\
+ 4f(x_6) + 2f(x_3) + 4f(x_7) + f(b) \left. \right] \frac{b-a}{8 \times 3} \] (13.73)

and recognize that \( (b-a)/8 = \Delta x \) is the width of a single strip when the interval \( a < x < b \) is divided into eight segments. Thus, we can write this last expression as
\[ 4T_8 - T_4 = \left[ f(a) + 4f(x_4) + 2f(x_2) + 4f(x_5) + 2f(x_1) \right. \\
+ 4f(x_6) + 2f(x_3) + 4f(x_7) + f(b) \left. \right] \left( \frac{\Delta x}{3} \right) \] (13.74)

which—\textit{mirabile dictu}—we recognize as Simpson’s rule for evaluating the integral with 8 divisions of the interval! We conclude that
\[ S_8 = \frac{4T_8 - T_4}{3} \] (13.75)

More generally, we could also conclude that
\[ S_{2n} = \frac{4T_{2n} - T_n}{3} \] (13.76)

The extrapolation formula of Eq. (13.76) applied to two successive evaluations by the trapezoidal rule gives the result of evaluation by Simpson’s rule!

A subroutine for trapezoidal integration can thus be used in a \textit{very} efficient algorithm for evaluating integrals by Simpson’s rule. We evaluate the integral twice by the trapezoidal rule for two values of \( n \), one of which is twice the other, finding \( T_n \) and \( T_{2n} \). Then we find the Simpson’s rule evaluation by exploiting the extrapolation formula of Eq. (13.76), which expresses the first step in what is called \textit{Romberg} integration. We do actual numerical integration \textit{only} with the trapezoidal rule, and we invoke the efficiency described in the first paragraph of this subsection in doing that. A routine for integration via the trapezoidal rule can thus be the workhorse for many other routines.

Indeed, Romberg integration goes beyond simply generating evaluations by Simpson’s rule from evaluations by the trapezoidal rule. Suppose we used the trapezoidal rule to generate the \textit{four} values \( T_n, T_{2n}, T_{4n}, \) and \( T_{8n} \). We could then use Eq. (13.76) to generate the \textit{three} values \( S_{2n}, S_{4n}, \) and \( S_{8n} \). In Romberg integration, we next generate a pair of still more accurate values from the formulae \( X_{4n} = (16S_{4n} - S_{2n})/15 \) and \( X_{8n} = (16S_{8n} - S_{4n})/15 \); and then we generate a further improved value from the formula \( Y_{8n} = (64X_{8n} - X_{4n})/63 \). This process can, of course, be continued indefinitely—though we rarely have to go even as far as we have described.

### 13.5.3 Assessing Error

Numerical evaluations, of course, only approximate the integral. Two distinctly different sorts of errors can occur. \textit{Truncation errors} arise because the integral has been approximated by a finite sum; \textit{roundoff errors} arise because computers do not store non-integers to 100% precision and, in the evaluation of a sum, the imprecision with which each component is represented within the computer
can accumulate as the number of arithmetic operations increases. Truncation errors become smaller as the width of strips is reduced. Roundoff errors, unfortunately, become more significant as the width of strips is reduced (because, with narrower strips, more arithmetic must be done). Usually, roundoff errors are negligible, the more so as the sophistication of the algorithm increases (and, hence, the amount of arithmetic decreases). Provided we do not strive for accuracy greater than about 1 part in $10^5$ or $10^6$ (with single precision floating point arithmetic), we can usually ignore roundoff errors. Thus, provided the function being integrated is such that the algorithm converges fairly rapidly with decreasing strip width, the quickest way to obtain a reasonably reliable estimate of truncation error is to evaluate the integral with two different step widths, the second being half of the first, and compare the two results. Presuming that roundoff error has not begun to be important, we can be confident that the second result is more accurate than the first. Thus, if the two agree to 1 part in $10^3$, say, we can with reasonable confidence, assume that the second value is good to one part in $10^3$. Indeed, the second value is probably better than that, but assessing its accuracy by this method would entail obtaining a third value by using a strip width half of that used to determine the second value.\footnote{We shall make this criterion a bit more explicit in the next paragraphs.} Indeed, one strategy for achieving a desired accuracy with reasonable certainty is to evaluate an integral repeatedly by a particular method, halving the strip width each time, and continuing until the new value received differs from its predecessor by less than the desired accuracy (though we must be careful not to push this approach so far that roundoff problems within the computer begin to become significant).\footnote{We shall see in later sections how we might decide when roundoff has started to be significant.}

From a more sophisticated perspective, numerical analysts have deduced expressions for the error in various Newton-Cotes formulae.\footnote{These formulae may be found in several sources, including Philip J. Davis and Philip Rabinowitz, 

*Numerical Integration* (Blaisdell Publishing Company, Waltham, MA, 1967), pp. 17, 20; and Milton Abramowitz and Irene A. Stegun, 

*Handbook of Mathematical Functions* (National Bureau of Standards Applied Mathematics Series 55, June, 1964), pp. 885ff.} For the midpoint formula of Eq. (13.61), for example,

$$
\left| \int_a^b f(x) \, dx - M_N \right| = \frac{(b - a)^3}{24N^2} \left| \frac{d^2 f}{dx^2} \right|_{x=\xi} \tag{13.77}
$$

where $\xi$ is some value of $x$ satisfying $a < \xi < b$—an expression that is valid provided the function $f(x)$ satisfies suitable requirements on continuity. The similar expressions

$$
\left| \int_a^b f(x) \, dx - T_N \right| = \frac{(b - a)^3}{12N^2} \left| \frac{d^2 f}{dx^2} \right|_{x=\xi} \tag{13.78}
$$

and

$$
\left| \int_a^b f(x) \, dx - S_N \right| = \frac{(b - a)^5}{180N^4} \left| \frac{d^4 f}{dx^4} \right|_{x=\xi} \tag{13.79}
$$

can be derived for the trapezoidal rule given by Eq. (13.62) and for Simpson’s rule given by Eq. (13.63), respectively. Again, $\xi$ is a value somewhere between $x = a$ and $x = b$, though it is not likely to have the same value in all three formulæ.

These results do not, of course, tell us how to determine the error exactly because they don’t tell us how to determine $\xi$ exactly. Even so, they are not entirely useless, having at least two particular values:

1. If it should happen in the first two cases that $d^2 f/dx^2 = 0$ or in the third case that $d^4 f/dx^4 = 0$ throughout the interval of integration, then the error is zero, since the right hand side of these expressions gives zero for all possible values of $\xi$. Thus, these formulæ confirm our previous assertions that the midpoint and trapezoidal rules will be 100% accurate for linear functions and that Simpson’s rule will be 100% accurate for cubic polynomials.
2. If $D_{\text{max}}(i)$ is the maximum value of $a^i f/dx^i$ in the interval $a < x < b$, then the above expressions support the inequalities

$$\left| \int_a^b f(x) \, dx - M_N \right| \leq \frac{(b-a)^3}{24N^2} D_{\text{max}}(2) \quad (13.80)$$

$$\left| \int_a^b f(x) \, dx - T_N \right| \leq \frac{(b-a)^3}{12N^2} D_{\text{max}}(2) \quad (13.81)$$

$$\left| \int_a^b f(x) \, dx - S_N \right| \leq \frac{(b-a)^5}{180N^4} D_{\text{max}}(4) \quad (13.82)$$

(though we must keep in mind that these approximations will frequently be extremely crude, so these upper bounds may well be very conservative). Provided that problems with computer roundoff do not begin to appear, we conclude from these results that an upper bound on the error in the midpoint and trapezoidal rules falls off like $1/N^2$ while that bound in Simpson’s rule falls off like $1/N^4$. Doubling $N$ therefore reduces the error in the midpoint and trapezoidal rules by a factor of four, while doubling $N$ reduces the error in Simpson’s rule by a factor of sixteen. With Simpson’s rule, every doubling of $N$ should gain at least one more decimal point in accuracy, so the convergence criterion described in the first paragraph in this subsection is particularly apt when Simpson’s rule is used.

### 13.5.4 Iterative and Adaptive Algorithms

In the previous subsections, we assumed that the user of a particular algorithm would actually view the value obtained for a succession of values of $N$ and decide personally when to stop by examining the changes that occur as $N$ is successively doubled. We can, of course, program a computer to make those decisions. One extremely common approach exploits Simpson’s rule (probably via the trapezoidal rule and Romberg integration) to obtain $S_2$, $S_4$, $S_8$, $\ldots$, compares each new value with its predecessor and stops when the absolute value of the difference is smaller than a tolerance—either absolute or relative—prescribed in advance. As a guard against an infinite loop, these algorithms should also stop if the desired tolerance has not been achieved in some maximum number of refinements and should print a warning when the desired tolerance has not been achieved. This method is said to be *iterative*, because it generates a succession of results, examines each new result in turn, and repeats the process until the new result meets or exceeds the prescribed tolerance. The points at which the function is evaluated, however, are determined ahead of time and are not influenced at all by the nature of the specific integrand to which the algorithm is applied.

Another family of algorithms (which may be iterative or noniterative) aims to minimize computational labor by estimating—though the methods for doing so are often crude—the accuracy obtained with each strip as the evaluation unfolds and shrinking or enlarging *that* strip to achieve a particular tolerance before going on to the next strip. In these *adaptive* methods, the points at which the function is evaluated are adjusted in response to the particular function being integrated. Because the assessment of accuracy at *a particular* strip can result either in shrinking or enlarging the width of *that*—or the next—strip, adaptive methods focus the computational effort in regions where the function varies rapidly and give less attention to regions in which the function varies slowly.

### 13.5.5 Gaussian Quadrature

The approach of *Gaussian quadrature* to numerical integration is more complicated than the Newton-Cotes approach but significantly better in some respects. In the Gaussian approach, both the points
at which the function is to be evaluated and the weights to be applied to each value are adjusted to achieve maximum accuracy when the function is approximated by a polynomial of a given order.

The development of a formula for Gaussian quadrature is simplified if we begin by introducing a set of \( m + 1 \) points \( t_i, (i = 0, 1, 2, \ldots, m) \) that divide the interval \( t_0 = a \leq t \leq b = t_m \) into \( m \) segments, the \( i \)-th of which extends over the interval \( t_{i-1} \leq t \leq t_i \). The values \( t_i \) may—but need not—be equally spaced. In this notation, we write the integral of interest as a sum of integrals over each segment, i.e., we write

\[
\int_a^b g(t) \, dt = \sum_{i=1}^m \int_{t_i}^{t_{i+1}} g(t) \, dt
\]  

(13.83)

To facilitate the discussion, however, we rescale and translate the variable in the \( i \)-th segment by

\[
x = \frac{2t - (t_{i+1} + t_i)}{t_{i+1} - t_i} \quad \text{or} \quad t = \frac{t_{i+1} + t_i}{2} + \frac{t_{i+1} - t_i}{2} x = t_i^{\text{mid}} + \frac{\Delta t_i}{2} x
\]  

(13.84)

where \( t_i^{\text{mid}} \) is the coordinate at the midpoint of the \( i \)-th segment and \( \Delta t_i \) is the width of the \( i \)-th segment. With this change, \( x \) ranges from \(-1\) to \(+1\) as \( t \) ranges from \( t_i \) to \( t_{i+1} \), so the integrals of interest now assume the form

\[
\int_a^b g(t) \, dt = \sum_{i=1}^m \int_{t_i}^{t_{i+1}} g(t) \, dt = \sum_{i=1}^m \frac{\Delta t_i}{2} \int_{-1}^1 g \left( t_i^{\text{mid}} + \frac{\Delta t_i}{2} x \right) \, dx = \sum_{i=1}^m \frac{\Delta t_i}{2} \int_{-1}^1 f_i(x) \, dx
\]  

(13.85)

where \( f_i(x) = g(t_i^{\text{mid}} + \Delta t_i x/2) \). In essence, then, we must evaluate an integral of the form

\[
\int_{-1}^1 f(x) \, dx
\]  

(13.86)

where we omit the subscript \( i \) on \( f \) for the sake of a simpler notation. If we can find a useful numerical evaluation for the integral in this standard form, then all else can be obtained by appropriate translations and rescalings.

The strategy for Gaussian integration now involves selecting the number of points—say \( N \)—at which the function is to be evaluated in the interval \(-1 < x < 1\), assuming an approximate formula of the form

\[
\int_{-1}^1 f(x) \, dx = \sum_{k=1}^N w_k f(x_k)
\]  

(13.87)

and then choosing both the weights \( w_k \) and the points of evaluation \( x_k \) to make this expression 100% accurate for a polynomial of as high an order as possible. Since we have \( 2N \) parameters to be determined, we should be able to make this expression accurate for a polynomial of order \( 2N - 1 \) with only \( N \) evaluations of the integrand.

To illustrate Gaussian integration more explicitly, let us derive a two-point formula, for which Eq. (13.87) would assume the more explicit form

\[
\int_{-1}^1 f(x) \, dx = w_1 f(x_1) + w_2 f(x_2)
\]  

(13.88)

We choose \( w_1, w_2, x_1 \) and \( x_2 \) so that the formula gives the correct answer for the special cases

\[
f(x) = 1, \text{ yielding that } \int_{-1}^1 dx = 2 = w_1 + w_2
\]  

(13.89)

\[
f(x) = x, \text{ yielding that } \int_{-1}^1 x \, dx = 0 = w_1 x_1 + w_2 x_2
\]  

(13.90)

\[
f(x) = x^2, \text{ yielding that } \int_{-1}^1 x^2 \, dx = \frac{2}{3} = w_1 x_1^2 + w_2 x_2^2
\]  

(13.91)

\[
f(x) = x^3, \text{ yielding that } \int_{-1}^1 x^3 \, dx = 0 = w_1 x_1^3 + w_2 x_2^3
\]  

(13.92)
Because the integral is a linear function of its integrand, a formula that yields the correct answer in these four cases will also yield the correct answer for any linear combination of these special cases, i.e., for any cubic polynomial. These four equations determine the four unknowns. Eqs. (13.90) and (13.92) imply that
\[ w_1 x_1 = -w_2 x_2 \quad \text{and} \quad w_1 x_1^3 = -w_2 x_2^3 \] (13.93)
which, when we divide the second by the first, yields \( x_1^2 = x_2^2 \), implying that \( x_1 = -x_2 \). (We reject the plus sign so the two values will be distinct.) Then Eqs. (13.91) and (13.89) yield that \( 2/3 = (w_1 + w_2) x_2^2 = 2 x_2^2 \ \Rightarrow \ x_2^2 = \frac{1}{3} \) (13.94)
from which we also conclude that \( x_2^3 = 1/3 \). Thus,
\[ x_1 = -\frac{1}{\sqrt{3}} \quad \text{and} \quad x_2 = \frac{1}{\sqrt{3}} \] (13.95)
Next, Eq. (13.90) implies that \( w_1 = w_2 \) and then Eq. (13.89) implies that \( w_1 = w_2 = 1 \). We conclude that
\[ \int_{-1}^{1} f(x) \, dx = f \left( -\frac{1}{\sqrt{3}} \right) + f \left( \frac{1}{\sqrt{3}} \right) \] (13.96)
The result is exact for cubic polynomials with only two evaluations of the function per strip! Simpson’s rule, also exact for cubic polynomials, requires three evaluations of the function per strip.\(^{14}\)

Returning to the original function and variables as laid out in Eq. (13.85), we finally find that Eq. (13.96) supports the expression
\[ \int_{a}^{b} g(t) \, dt = \sum_{i=1}^{m} \Delta t_i \left[ g \left( t_{i}^{\text{mid}} - \frac{\Delta t_i}{2} \frac{1}{\sqrt{3}} \right) + g \left( t_{i}^{\text{mid}} + \frac{\Delta t_i}{2} \frac{1}{\sqrt{3}} \right) \right] \]
\[ = \sum_{i=1}^{m} \frac{\Delta t_i}{2} \sum_{j=1}^{2} w_j g \left( t_{i}^{\text{mid}} + \frac{\Delta t_i}{2} x_j \right) \] (13.97)
The strategy invoked to develop the two-point Gaussian formula can also be applied to deduce higher order formulae. For a three-point formula, for example, we would have three points and three weights, and we would expect to be able to choose these unknowns to generate a formula that would be exact for a fifth-degree polynomial. The five-point formula
\[ \int_{-1}^{1} f(x) \, dx = 0.23692689 f(-0.90617985) + 0.47862867 f(-0.53846931) \]
\[ + 0.56888889 f(0.00000000) \]
\[ + 0.47862867 f(0.53846931) + 0.23692689 f(0.90617985) \] (13.98)
which is 100% accurate for polynomials of the ninth-degree or lower, is among the most popular of the formulae in this class. For the sake of later examples, we note that, in this expression
\[ w_1 = 0.23692689 \quad x_1 = -0.90617985 \]
\[ w_2 = 0.47862867 \quad x_2 = -0.53846931 \]
\[ w_3 = 0.56888889 \quad x_3 = +0.00000000 \] (13.99)
\[ w_4 = 0.47862867 \quad x_4 = +0.53846931 \]
\[ w_5 = 0.23692689 \quad x_5 = +0.90617985 \]

\(^{14}\) The advantage isn’t that great, however, because, for Simpson’s rule, the upper evaluation for one strip could also be used as the lower evaluation for the next strip. No such feature applies to the two-point—or to any—Gaussian integration formula. The advantage of the Gaussian approach increases, however, as \( N \) in Eq. (13.87) increases.
Further, returning to the original variable, we note that

\[
\int_a^b g(t) \, dt = \sum_{i=1}^m \frac{\Delta t_i}{2} \sum_{j=1}^5 w_j g \left( t_{\text{mid}}^i + \frac{\Delta t_i}{2} x_j \right)
\]  
\tag{13.100}

As an aside, note that the points \( x_1 \) and \( x_2 \) at which we have evaluated the function for two-point Gaussian quadrature are the two roots of the second Legendre polynomial, \( L_2(x) = \frac{1}{2} (3x^2 - 1) \) and the weight to be applied to \( f(x_i) \) is given by \((2/[(1-x^2)(dL_2(x_i)/dx)]\). More generally, for an \( N \)-point Gaussian integration, we would discover that

\[
L_N(x_i) = 0 \text{ and } w_i = \frac{2}{(1-x^2)(dL_N(x_i)/dx)^2}
\]  
\tag{13.101}

and that

\[
\left| \int_{-1}^1 f(x) \, dx - \sum_{k=1}^N w_k f(x_k) \right| = \frac{2^{2N+1}(N!)^4}{(2N+1)[(2N)!]^3} \left| \frac{d^{2N} f}{dx^{2N}} \right|_{x=\xi}
\]  
\tag{13.102}

where \(-1 < \xi < 1\). This result shows that the \( N \)-point formula of this type will be 100% accurate for polynomials of degree \( 2N - 1 \) or lower—a property which we have already inferred informally. Because of the role played by the Legendre polynomials in these formulae, they are sometimes referred to as Gauss-Legendre formulae.\(^{16}\)

### 13.8 Evaluating Integrals Numerically with OCTAVE

Note: All OCTAVE program (.m) files referred to in this chapter are available in the directory \$HEAD/octave, where (as defined in the Local Guide) \$HEAD must be replaced by the appropriate path for your site. At some sites, this directory or some other directory containing these files may also have been placed in OCTAVE’s default search path. If so, the files can be found by OCTAVE without explicit specification of a path. Otherwise, you will have to use the full path to copy them into your default directory to access them.

One-dimensional integrals can be evaluated numerically either by using OCTAVE’s elementary commands as described in Section 13.8.1 or, more simply, by invoking one of the built-in routines as described in Section 13.8.2.

#### 13.8.1 Using Elementary Commands

Relatively simple sequences of elementary commands can implement one or another of the algorithms described in Section 13.5. If, for example, we seek an evaluation of the integral

\[
I = \text{erf}(1) = \frac{2}{\sqrt{\pi}} \int_0^1 e^{-x^2} \, dx
\]  
\tag{13.103}

which appears in the expression of Eq. (13.41) for the probability of finding a quantum oscillator in the classically forbidden region, we might invoke the trapezoidal rule as presented in Eq. (13.62), divide the interval \( 0 \leq x \leq 1 \) into \( n = 4 \) segments, and use the statements\(^{17}\)

\[
y = 0.5*\text{f}(1) + \sum (\text{f}(2:n) + 0.5*\text{f}(n+1))
\]


\(^{16}\)A more general integral that can be approached with the techniques of this subsection has the form \( \int_a^b f(x) \, w(x) \, dx \), where \( w(x) \) is a weighting function. In the case we dealt with, \( w(x) = 1 \) and, in our rescaling, the interval became the interval from \(-1\) to \(1\). That the Legendre polynomials \( L_i(x) \) are orthogonal on the interval \(-1 < x < 1\) with weight \( w(x) = 1 \) is part of the reason that the roots of these polynomials ultimately emerged as important. For other weight functions and other intervals, a different set of polynomials would have played the role of the Legendre polynomials. Thus, there are several different types of Gaussian quadrature, each specific to a particular weight and basic interval.

\(^{17}\)Alternatively, as long as \( n \geq 2 \), the sixth, seventh, and eighth statements in this code can be replaced with the single statement \( y = 0.5*\text{f}(1) + \sum(\text{f}(2:n)) + 0.5*\text{f}(n+1)\).
Table 13.1: Values of erf(1.0) obtained by the trapezoidal rule and a user-constructed program in OCTAVE. Values were determined by double-precision calculations and all resulting digits are shown, even though not all are significant.

<table>
<thead>
<tr>
<th>n</th>
<th>(T_n)</th>
<th>n</th>
<th>(T_n)</th>
<th>n</th>
<th>(T_n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.77174332258054</td>
<td>64</td>
<td>0.84268390204498</td>
<td>4096</td>
<td>0.842700788825994</td>
</tr>
<tr>
<td>2</td>
<td>0.825262955596749</td>
<td>128</td>
<td>0.842696570249296</td>
<td>8192</td>
<td>0.8427007919873</td>
</tr>
<tr>
<td>4</td>
<td>0.88367777441205</td>
<td>256</td>
<td>0.84269737276221</td>
<td>16384</td>
<td>0.842700792918783</td>
</tr>
<tr>
<td>8</td>
<td>0.841619221244768</td>
<td>512</td>
<td>0.842700529031443</td>
<td>32768</td>
<td>0.842700792885281</td>
</tr>
<tr>
<td>16</td>
<td>0.842430505490233</td>
<td>1024</td>
<td>0.842700726970153</td>
<td>65536</td>
<td>0.842700792945679</td>
</tr>
<tr>
<td>32</td>
<td>0.842633227681257</td>
<td>2048</td>
<td>0.842700776454825</td>
<td>131072</td>
<td>0.842700792933611</td>
</tr>
</tbody>
</table>

```
>> format long
>> n = 4;
>> dx = 1.0/n;
>> x = [ 0.0 : dx : 1.0 ];
>> f = 2.0*exp(-x.^2)/sqrt(pi);
>> y = 0.5*f(1);
>> for i = 2:n y=y+f(i); end
>> y = y + 0.5*f(n+1);
>> dx*y
ans = 0.838367777441205
```

where, of course, not all of the digits are significant. To illustrate the reduction of truncation error with decreasing step size, we re-execute this procedure, starting with \(n = 1\) and successively doubling \(n\), finding the values shown in Table 13.1. On the basis of this succession of numbers, however, we conclude that, to six decimal places, the value of the integral is 0.842701. The trapezoidal rule took 513 evaluations of the function to reach that value.

Note in this example that the result for \(n = 1\) is quite inaccurate. As \(n\) increases, however, the result becomes more and more accurate. After a time (here at about \(n = 512\)) the result stabilizes—at least to the first six or seven digits after the decimal point—and further increase of \(n\) makes almost no difference in the value to that precision. A graph of \(T_n\) versus \(n\) would start low, increase to this stable value and remain horizontal as \(n\) is further increased. Indeed, in this example, that graph would remain horizontal all the way to \(n = 131072\). Were we to increase \(n\) still further, we would sooner or later find that this graph would begin to depart from the value at which it stabilized, at first slowly but then more dramatically. The point at which that departure begins is the point at which roundoff errors begin to become significant. Fortunately, \(n\) exhibits a substantial range of values that are simultaneously large enough to keep truncation error at bay and small enough to prevent significant roundoff error. Short of changing to a more sophisticated algorithm, the most accurate value we can obtain is the value at which further increase of \(n\) (for awhile) makes little difference in the value obtained.

We illustrate with three alternative and more efficient approaches. First, note that Richardson extrapolation applied to \(T_1\) and \(T_2\) and then again to \(T_2\) and \(T_4\) yields the values

\[
S_2 = \frac{4 \cdot T_2 - T_1}{3} = \frac{4 \cdot 0.825262955596749 - 0.77174332258054}{3} = 0.843102830042981 \quad (13.104)
\]

and

\[
S_4 = \frac{4 \cdot T_4 - T_2}{3} = \frac{4 \cdot 0.838367777441205 - 0.825262955596749}{3} = 0.842736015389357 \quad (13.105)
\]

Continuing this process of Richardson extrapolation, we find the values \(S_8 = 0.842703035845956\), \(S_{16} = 0.842700933572055\), and \(S_{32} = 0.842700801744932\). Because \(S_{16}\) and \(S_{32}\) agree to six digits,
we conclude that \( I = 0.842701 \) to the sixth decimal place. We have here arrived at the same conclusion as in the previous paragraph, but only 17 evaluations of the function were necessary. The trapezoidal rule coupled with Richardson extrapolation is clearly more efficient than the trapezoidal rule alone, at least with this integral.

Second, if we instead adopt Simpson’s rule as in Eq. (13.63), we would need to modify the above procedure to recognize that each evaluation of the function requires a different weight, i.e., the first and last values of the function must be multiplied by 1, the second value and alternate values thereafter must be multiplied by 4, and the third value and alternate values thereafter must be multiplied by 2. To achieve this end, we create a vector of weights having the value \([1,4,2,4,\ldots,2,4,1]\). We might invoke the statements\(^\text{18}\)

\[
\begin{align*}
\text{\texttt{>> n = 2;}} & \quad \text{Set number of segments.} \\
\text{\texttt{>> dx = 1.0/n;}} & \quad \text{Set step size.} \\
\text{\texttt{>> x = [0.0 : dx : 1.0];}} & \quad \text{Set values of } x. \\
\text{\texttt{>> f = 2.0*exp(-x.^2)/sqrt(pi);}} & \quad \text{Evaluate integrand.} \\
\text{\texttt{>> clear w}} & \quad \text{Set all weights to 0.} \\
\text{\texttt{>> w(1) = 1.0; w(n+1) = 1.0;}} & \quad \text{Set first and last weights to 1.} \\
\text{\texttt{>> for i=2:n w(i)=4.0; end}} & \quad \text{Set proper weights to 4.} \\
\text{\texttt{>> for i=3:2:n-1 w(i) = 2.0; end}} & \quad \text{Set remaining weights to 2.} \\
\text{\texttt{>> y = 0.0;}} & \quad \text{Initialize accumulator.} \\
\text{\texttt{>> for i=1:n+1 y = y + w(i)*f(i);}} & \quad \text{Add each strip with proper weight, suppressing output.} \\
\text{\texttt{end}} & \\
\text{\texttt{>> dx*y/3.0}} & \quad \text{Rescale; display result.}
\end{align*}
\]

where, again, not all digits are significant. If this process is repeated with \( n = 4 \), the end result is 0.842736051389357. With \( n = 32 \) and \( n = 64 \), this process yields the values 0.842700801744932 and 0.842700793499513, respectively. Reassuringly, all of these values agree with the values obtained by applying Richardson extrapolation to the values obtained in the previous paragraph with the trapezoidal rule. We again conclude that, to six digits, \( I = 0.842701 \).

Third, we could adopt the Gaussian approach. If, for example, we identify \( t \) in Eq. (13.97) with \( x \) in Eq. (13.103) and \( g(t) \) with \( 2e^{-x^2}/\sqrt{\pi} \) and we divide the interval \( 0 \leq x \leq 1 \) into \( m = 4 \) segments of equal width, we might invoke the OCTAVE statements

\[
\begin{align*}
\text{\texttt{>> format long}} & \quad \text{Set output format.} \\
\text{\texttt{>> pts = [-1.0/sqrt(3.0) 1.0/sqrt(3.0)];}} & \quad \text{Set evaluation points } x_1 \text{ and } x_2. \\
\text{\texttt{>> w = [1.0 1.0];}} & \quad \text{Set weights } w_1 \text{ and } w_2. \\
\text{\texttt{>> a = 0.0; b = 1.0;}} & \quad \text{Set limits.}
\end{align*}
\]

to set the evaluation points, weights, and limits. Then we would execute the statements

\[
\begin{align*}
\text{\texttt{>> m = 4;}} & \quad \text{Set number of segments.} \\
\text{\texttt{>> dx = (b-a)/m;}} & \quad \text{Set width of each segment.} \\
\text{\texttt{>> x = [a : dx : b];}} & \quad \text{Set values dividing segments.} \\
\text{\texttt{>> for i=1:m xmid(i)=(x(i)+x(i+1))/2.0; end}} & \quad \text{Evaluate midpoints.}
\end{align*}
\]

to set several parameters and determine the coordinates at the midpoints of the segments. Finally, we evaluate the sum, multiply with an overall factor whose inclusion was postponed, and display the result with the statements

\[
\begin{align*}
\text{\texttt{\ldots}} & \\
\text{\texttt{ans = 0.843102830042981}} & \quad \text{Rescale; display result.}
\end{align*}
\]

\(^{18}\)Alternatively, the sum of the products of the components of \( w \) and \( f \) could be viewed as a multicomponent vector dot product and evaluated with the single statement \( y = f \cdot \text{transpose}(w) \).
13.8. EVALUATING INTEGRALS NUMERICALLY WITH OCTAVE

```octave
>> int = 0.0;
>> for i=1:m
    fcts = 0.0;
    for j = 1:2 fcts = fcts + w(j)*exp(-(xmid(i)+dx*pts(j)/2.0)^2); end;
    int = int + fcts; end
>> int = dx*int/sqrt(pi)
ans = 0.842699298101774
```

finding that the two-point Gaussian formula with two equal divisions of the interval of integration yields a result that is correct to five digits. With 1, 2, 4, and 8 divisions, the results are 0.842441892522547, 0.842677323862959, 0.842699298101774, and 0.842700699207473, respectively—and we find that the two-point Gaussian formula yields a result correct to four digits even with only two divisions of the interval of integration and correct to six digits with eight divisions.

The coding worked out in the previous paragraph is easily adapted to express the five-point Gaussian formula in Eq. (13.99). We begin by invoking the statements

```octave
>> format long
>> pts = [-0.90617985 -0.53846931 0.0 0.53846931 0.90617985 ];
>> w = [ 0.23692689 0.47862867 0.56888889 0.47862867 0.23692689 ];
>> a = 0.0; b = 1.0;
```

to set the evaluation points \(x_1, \ldots, x_5\), the weights \(w_1, \ldots, w_5\) and the limits. Then, we prepare to evaluate the integral with the statements

```octave
>> m = 4;  \text{Set number of segments.}
>> dx = (b-a)/m;
>> x = [a : dx : b];  \text{Set width of each segment.}
>> for i=1:m xmid(i)=(x(i)+x(i+1))/2.0; end;  \text{Set values dividing segments.}
```

Finally, we evaluate the sum, multiply with an overall factor whose inclusion was postponed, and display the result with the statements

```octave
>> int = 0.0;
>> for i=1:m
    fcts = 0.0;
    for j = 1:5 fcts = fcts + w(j)*exp(-(xmid(i)+dx*pts(j)/2.0)^2); end;
    int = int + fcts; end
>> int = dx*int/sqrt(pi)
ans = 0.842700797137631
```

finding that the five-point Gaussian formula with four equal divisions of the interval of integration yields a result that is correct to six digits. Even more amazing, this five-point formula gives the result 0.8427007899282891—which agrees with the above result to seven digits—with \(m = 1\), i.e., when the entire interval of integration is treated as a single segment!

13.8.2 Built-In Integration Routines

One-dimensional integrals can be evaluated numerically either by using OCTAVE's elementary commands or, more simply, by invoking one of the built-in routines, specifically,

- `trapz` and `cumtrapz`, which use the trapezoidal rule.
Table 13.2: The OCTAVE function gausint.m.

```octave
function y = gausint(t)
# GAUSINT - Defines Gaussian lineshape
# GAUSINT defines the integrand for evaluating the error function.

tmp1 = 2.0/sqrt(pi);
y = tmp1*exp(-t.^2);
endfunction
```

- `quadv`, which uses a recursive, adaptive vectorized Simpson's rule.
- `quadl`, which uses a recursive, adaptive Lobatto rule.
- `quad`, which uses Gaussian quadrature.
- `quadgk`, which uses an adaptive Gauss-Konrod rule.
- `quadcc`, which uses doubly-adaptive Clenshaw-Curtis rules.

The optimal method to use depends on the character of the integrand. For smooth integrands, all methods will work, though some will require more calculation than others to achieve a particular accuracy. If the integrand is smooth and the integrand is oscillatory and/or the integral has infinite limits, `quadgk` and `quadcc` should be used. For integrands with singularities in the interval of integration, `quadcc` should be chosen. In the following sections, we illustrate the use of only a few of these options. Full details on each of these commands can be found in the OCTAVE manuals.

The simplest of these routines—`trapz`—is straightforward and non-adaptive. It simply uses the trapezoidal rule to evaluate \( \int y \, dx \) when supplied with two vectors \( x \) and \( y \), the first containing equally spaced values of the independent variable and the second containing values of the dependent variable at the points in \( x \). Thus, for example, we might evaluate the integral in the previous section via the trapezoidal rule with 16 segments with the statements

```octave
>> format long
>> x = linspace( 0.0, 1.0, 17 );
>> y = 2.0*exp(-x.^2)/sqrt(pi);
>> I = trapz( x, y )
I = 0.842430505490233
```

We have, of course, obtained \( T_{16} \)—and this result agrees with previous evaluations of the integral by this method with 16 divisions. The routine `cumtrapz` differs from `trapz` simply by printing out the accumulating value of the integral as each next step is added to the result, and it may produce a substantial volume of output.

The second method—`quadv`—uses Simpson's rule, and the third—`quadl`—uses adaptive Lobatto quadrature. Before invoking either of these routines, we must create a function M-file defining the integrand as a function of a single scalar argument (the integration variable).\(^{19}\) The construction of such files is described in Section 4.9. For the integral in Eq. (13.41), for example, the function M-file might be as listed in Table 13.2.

Having stored the M-file in the user's default directory with a name identical to that of the function name and with the file extension .m (here `gausint.m`), we would evaluate the integral and display its value with a statement having the general form\(^{20}\)

\(^{19}\)Inclusion of parameters in the definition of the function will be discussed in Section 13.8.8.

\(^{20}\)The first argument in these commands can also be the function name enclosed in single or double quotation marks, e.g. `q = quadv( 'function_name', ... )`, or a handle identifying the function. See footnote 36 in Chapter 4.
13.8. EVALUATING INTEGRALS NUMERICALLY WITH OCTAVE

\[ q = \text{quadv}( \text{@function\_name}, a, b, tol ) \]
\[ [q, n] = \text{quadv}( \text{@function\_name}, a, b, tol ) \]
\[ q = \text{quadl}( \text{@function\_name}, a, b, tol ) \]

Here, \text{function\_name}, which is mandatory, identifies the file in which the integrand is defined; \text{a} and \text{b}—also mandatory—are the lower and upper limits of integration; and \text{tol}, which is optional, is a single number specifying the desired absolute accuracy (default $10^{-6}$ for \text{quadv} and the value of the system variable \text{eps}, which is $2.2204 \times 10^{-16}$ by default, for \text{quadl}). As here illustrated, these routines return the value of the integral in the specified variable, here \text{q}. If the second form for the output from \text{quadv} is used, \text{quadv} also returns the number of times the integrand is evaluated in arriving at the result. The routine \text{quadl} does not offer that option.

For the example of Eq. (13.41), the integral is evaluated with the absolute accuracies $10^{-3}$, $10^{-6}$ (the default) and $10^{-9}$ and displayed by the statements

```matlab
>> format long
>> [ q, n ] = quadv( @gausint, 0.0, 1.0, 1.0e-3 )
q = 0.842703035845956
n = 9
>> [ q, n ] = quadv( @gausint, 0.0, 1.0 )
q = 0.842700801744932
n = 33
>> [ q, n ] = quadv( @gausint, 0.0, 1.0, 1.0e-9 )
q = 0.842700792973848
n = 117
```

Because of the tolerance set, we have no reason to believe that the first value is good beyond the third digit. The second value should be good to the sixth digit, i.e., $I = 0.842701$, which certainly agrees with the values obtained earlier. The second and third values certainly agree to six digits, though we might expect the third value to be accurate to eight or nine digits. In other words, if we believe the tolerance specified in the third statement, we might presume the value of the integral to be $0.842700793$ to nine digits.

The syntax of statements invoking \text{quadl} is identical to that described above for \text{quadv}. Thus, the statements

```matlab
>> q = quadl( @gausint, 0.0, 1.0, 1.0e-3 )
q = 0.842700792949715
>> q = quadl( @gausint, 0.0, 1.0, 1.0e-6 )
q = 0.842700792949715
>> q = quadl( @gausint, 0.0, 1.0, 1.0e-9 )
q = 0.842700792949715
```

use \text{quadl} to the evaluate the integral to the tolerances $10^{-3}$, $10^{-6}$ and $10^{-9}$. Curiously, \text{quadl} returns the same value for all three tolerances.

At base, the syntax for \text{quad} is the same as that for \text{quadv} and \text{quadl}, except that the optional argument \text{tol} is a two-component vector whose first component is the desired absolute tolerance and whose second component is the desired relative tolerance, both of which default to $\sqrt{\text{eps}}$. More specifically,
Table 13.3: The OCTAVE function \texttt{moment.m}.

\begin{verbatim}
function y = moment( lambda )
  # MOMENT - Defines integrand for moment of inertia
  # MOMENT defines the integrand for evaluating the moment of
  # inertia of a semicircular plate.
  tmp = lambda.^2;
  y = 4.0*tmp.*sqrt(1.0-tmp)/pi;
endfunction
\end{verbatim}

\verbatim
q = quad( @function_name, a, b, tol )
[ q, ecode, n, err ] = quad( @function_name, a, b, tol )
\end{verbatim}

where the symbols are as defined for \texttt{quadv}. In addition, in the second form for the output, \texttt{ecode} is a code that indicates the nature of the success of the integration (0 = success) and \texttt{err} is an estimate of the actual error in the value of the integral. The statements

\begin{verbatim}
>> [ q, ecode, n, err] = quad( @gausint, 0.0, 1.0, [1.0e-3, 0.0] )
q = 0.842700792949715 [ecode = 0, n = 21, err = 9.3...e-015]
>> q = quad( @gausint, 0.0, 1.0, [1.0e-6, 0.0] )
q = 0.842700792949715 [ecode = 0, n = 21, err = 9.3...e-015]
>> q = quad( @gausint, 0.0, 1.0, [1.0e-9, 0.0] )
q = 0.842700792949715 [ecode = 0, n = 21, err = 9.3...e-015]
\end{verbatim}

will evaluate the sample integral with \texttt{quad}. The efficiency of Gaussian integration is evident in these results.

13.8.3 Moment of Inertia

To evaluate the integral appearing in Eq. (13.16) for the moment of inertia of a semicircular plate, we must first create the function M-file listed in Table 13.3 and store it in a file named \texttt{moment.m}. Then, the simple statements

\begin{verbatim}
>> format long
>> [ I, n ] = quadv( @moment, 0.0, 1.0, 1.0e-3 )
I = 0.249694928749505
n = 21

>> [ I, n ] = quadv( @moment, 0.0, 1.0, 1.0e-6 )
I = 0.24999538224446
n = 77

>> I = quadl( @moment, 0.0, 1.0, 1.0e-3 )
I = 0.249914272120002
\end{verbatim}
Table 13.4: Dependence of trapezoidal result on number of steps.

<table>
<thead>
<tr>
<th>( n )</th>
<th>( I )</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.249626708043829</td>
</tr>
<tr>
<td>1000</td>
<td>0.249988166015548</td>
</tr>
<tr>
<td>10000</td>
<td>0.249999625683606</td>
</tr>
<tr>
<td>100000</td>
<td>0.249999988162782</td>
</tr>
<tr>
<td>1000000</td>
<td>0.249999999625671</td>
</tr>
</tbody>
</table>

\[
\text{\texttt{quadl( @moment, 0.0, 1.0, 1.0e-6 )}}
\]
\[
\text{I = 0.249999933894595}
\]

evaluate the integral to one part in 10\(^3\) and one part in 10\(^6\) with \texttt{quad} and \texttt{quadl}. These results are clearly in agreement with one another and with those obtained by other methods in this chapter, provided we don’t believe more than five or six digits. Rounded to five digits, the second and fourth of these results are 0.25000.

If we didn’t know, however, what the “exact” value should be, we would have to be more careful about interpreting the above numerical result. We could, for example, exploit the simpler, non-adaptive command \texttt{trapz} with the statements

\[
\text{\texttt{\texttt{>> n = 10;}}}
\]
\[
\text{\texttt{\texttt{>> x = linspace( 0.0, 1.0, n+1 );}}}
\]
\[
\text{\texttt{\texttt{>> y = 4.0 * x.^2 .* sqrt(1.0-x.^2)/pi;}}}
\]
\[
\text{\texttt{\texttt{>> I = trapz( x, y )}}}
\]
\[
\text{\texttt{\texttt{\texttt{>> I = 0.238496332979749}}}}
\]
\[
\text{\texttt{\texttt{>> format short}}}
\]

to evaluate the desired integral via the trapezoidal rule with 10 segments in the interval \(0 \leq x \leq 1\). We can repeat this process for larger values of \(n\), finding the values shown in Table 13.4. Since, as discussed in Section 13.5.3, the error in the trapezoidal rule decreases in inverse proportion to the square of \(n\), each new step in this sequence should yield an improvement of a factor of 100 (two decimal digits) in precision. By replacing the automatic criterion with a personal examination, we can be more confident that the value of this integral is approaching 0.25000.\(^{21}\)

### 13.8.4 Quantum Probabilities

In Section 13.8.2, we have already evaluated the integral that appears in the determination of the probability that a quantum harmonic oscillator in its ground state will be found outside the classical turning point. According to Eq. (13.41), that probability is given by the expression

\[
P(|x| > |x_{\text{turn}}|) = 1.0 - \text{erf}(1.0) = 1.0 - 0.842701 = 0.157299
\]

We conclude that, in a bit over 15\% of the measurements, the quantum oscillator will be found in the classically forbidden region!

\(^{21}\)Remember that OCTAVE does its internal arithmetic in double precision, so we have probably not yet begun to be affected by problems from internal roundoff.
13.8.5 Integrals as Functions of the Upper Limit

To evaluate an integral as a function of a parameter, the routine used must be invoked repeatedly in a loop. With each execution of the loop, the parameter assumes a new value and the current values of the parameter and of the corresponding integral are stored for later examination. This process is easiest to implement when the parameter is the upper limit of the integral as, for example, in the integral

\[ g(x) = \int_0^x f(t) \, dt \]  \hspace{1cm} (13.107)

We begin by constructing a function M-file to define the integrand \( f(t) \), say `integ.m`. Then, supposing that we want to evaluate the integral as a function of its upper limit \( x \) for \( a \leq x \leq b \), we invoke the OCTAVE statements\(^{22}\)

\[
\begin{align*}
N &= \langle \text{appropriate value} \rangle; \hspace{2cm} \text{Set number of segments in interval.} \\
x &= \text{linspace} (a, b, N+1); \hspace{2cm} \text{Set values for upper limit.} \\
\text{for } i = 1:N+1 & \hspace{2cm} \text{Evaluate integral for each upper limit in } x, \\
\quad g(i) &= \text{Routine} ( \text{@integ}, 0, x(i) ); \hspace{2cm} \text{using a chosen routine and accepting the default tolerance.} \\
\text{end}; & \hspace{2cm} \\
\end{align*}
\]

Upon execution of these statements, values of \( g(x) \) as a function of \( x \) for the selected values of \( x \) will be stored in the vector \( g \) and can be further processed as desired. In particular, the statement \( \text{plot}(x,g) \) — or some embellishment thereof — will generate a graph of \( g \) versus \( x \).\(^{23}\)

13.8.6 The Error Function

To evaluate the error function as given by Eq. (13.26), we begin by constructing the M-file giving the integrand. The function `gausint.m` defined in Section 13.8.2 can be used here also. Thus, supposing we want \( x \) to range from 0.0 to 3.0 in steps of 0.1, we invoke the OCTAVE statements

\[
\begin{align*}
\text{>> } N &= 30; & \hspace{2cm} \text{Set graphics toolkit.} \\
\text{>> } x &= \text{linspace} (0.0, 3.0, N+1); & \hspace{2cm} \text{Set number of segments.} \\
\text{>> for } i = 1:N+1 & \hspace{2cm} \text{Specify values of upper limit.} \\
\quad \text{erf}(i) &= \text{quad} ( \text{@gausint}, 0, x(i) ); & \hspace{2cm} \text{Evaluate integral for each upper limit in } x, \\
\text{end}; & \hspace{2cm} \text{accepting defaults for keywords.} \\
\text{>> plot}(x, \text{erf}, 'linewidth', 4, ... & \hspace{2cm} \text{Plot error function.} \\
\quad 'color', 'black') & \hspace{2cm} \\
\text{>> title(‘Error Function’, ’fontsize’, 20 )} & \hspace{2cm} \text{Title graph.} \\
\text{>> xlabel( ‘x’, ’fontsize’, 16 )} & \hspace{2cm} \text{Label axes.} \\
\text{>> ylabel( ‘erf(x)’, ’fontsize’, 16 )} & \hspace{2cm} \text{Turn on grid.} \\
\text{>> grid on} & \hspace{2cm} \text{Adjust size for labels} \\
\text{>> set( gca, ’fontsize’, 16 )} & \hspace{2cm} \\
\end{align*}
\]

The resulting graph is shown in Fig. 13.10.

13.8.7 The Cornu Spiral

To determine the Cornu spiral, we use `quad` to generate vectors \( C \) and \( S \) containing values of the defining integrals as given in Eq. (13.27) over a suitable range of upper limits and then plot the values in \( S \) versus the values in \( C \). First we create the M-files in Table 13.5 to provide the two

\^22\We might, of course, use `quadl` instead of `quad` in this example.
\^23\This sequence of statements actually is computationally inefficient. We might increase the efficiency by recognizing, for example, that \( g(x + \Delta x) = g(x) + \int_x^{x+\Delta x} f(t) \, dt \) and obtain integrals for larger \( x \) by adding an appropriate increment to already evaluated integrals for smaller \( x \). For our present purposes, that approach unnecessarily complicates the algorithm of evaluation.
13.8. EVALUATING INTEGRALS NUMERICALLY WITH OCTAVE

Figure 13.10: The error function.

Table 13.5: The OCTAVE functions spiralc.m and spirals.m.

```octave
function y = spiralc(t)
    # SPIRALC - returns cosine integrand for Cornu spiral.
    # SPIRALC defines the integrand for the integral giving the
    # vertical coordinate of the Cornu spiral.
    y = cos( pi * t.^2 / 2.0 );
endfunction

function y = spirals(t)
    # SPIRALS - returns sine integrand for Cornu spiral.
    # SPIRALS defines the integrand for the integral giving the
    # horizontal coordinate of the Cornu spiral.
    y = sin( pi * t.^2 / 2.0 );
endfunction
```

integrands, storing them in the user’s default directory with the names spiralc.m and spirals.m. For definiteness, we elect to explore the spiral over the range $-5 \leq u \leq 5$, though we evaluate only integrals for $0 \leq u \leq 5$ and obtain values for $-5 \leq u \leq 0$ by recognizing that both integrals are odd functions of $u$. First, we generate the necessary values of the upper limit and evaluate the integrals with the statements
>> N = 100;
>> u = linspace( 0, 5.0, N+1 );
>> for i = 1:N+1
    CP(i)=quad( @spiralc, 0, u(i) );
    SP(i)=quad( @spirals, 0, u(i) );
    CN(N + 2 - i) = -CP(i);
    SN(N + 2 - i) = -SP(i);
end

Finally, we assemble the variables to be plotted and plot the graph with the statements

>> C = [CN, CP];
>> S = [SN, SP];
>> plot( C, S, 'linewidth', 2, ... 'color', 'black' )
>> axis( [-1.0 1.0 -1.0 1.0 ] )
>> axis equal
>> grid on
>> title( 'Cornu Spiral', 'fontsize', 20 )
>> xlabel( 'C(u)', 'fontsize', 16 )
>> ylabel( 'S(u)', 'fontsize', 16 )
>> set( gca, 'fontsize', 16 )

The resulting graph is shown in Fig. 13.11.

13.8.8 Integrals as Functions of an Internal Parameter

The situation in which an integral of interest is a function of a parameter in the integrand is more difficult because we must somehow sneak the parameters in the integrand through the integration

---

24This sequence of statements actually has two glitches. First, it is computationally inefficient in the way described in footnote 23. Second, and less significantly, the values of S(0) and C(0) appear twice in the vectors finally plotted.
13.8. EVALUATING INTEGRALS NUMERICALLY WITH OCTAVE

routine and into the function called by those routines to evaluate the integrand. One technique for achieving this objective is to exploit global variables as discussed in Section 4.9 and in Section 11.9.4. Using global variables will work with all of OCTAVE’s integration routines. Conveniently, however, \texttt{quadv} and \texttt{quadl} offer an alternative. Calls to these routines can take advantage of the existence of additional arguments. The full structure of statements involving these commands is

\begin{verbatim}
q = quadv( @function_name, a, b, tol, trace, p1, p2, ... )
[ q, n ] = quadv( @function_name, a, b, tol, trace, p1, p2, ... )
q = quadl( @function_name, a, b, tol, trace, p1, p2, ... )
\end{verbatim}

Here—beyond the quantities already defined—a non-zero value\(^{25}\) for \texttt{trace} will display the iterative, recursive progress towards the final result as the integration proceeds,\(^{26}\) and \texttt{p1}, \texttt{p2}, \ldots are parameters that will be passed by \texttt{quadv} or \texttt{quadl} to the function defining the integrand, provided that the function is defined with those parameters as additional arguments. To accept the default values of \texttt{tol} and/or \texttt{trace} while specifying parameters, simply set either/both to the empty matrix \([\ ]\).

To invoke this feature, we must do two things. First, we must define the integrand with the parameters as additional arguments to the function. The appropriate file would then have the general format

\begin{verbatim}
function y = FunctionName( x, p1, p2, p3, ... )
  # Explanatory comments.
  Statements to evaluate integrand, using p1, p2, p3, ... for the parameters.
  y = integrand;
endfunction
\end{verbatim}

The quantities \texttt{FunctionName}, \texttt{x}, and \texttt{integrand} have the same meaning as before, and the file must be stored with the name \texttt{FunctionName.m}.

Second, having defined the function M-file in this new way and stored it in the default directory, we invoke OCTAVE to evaluate and display the desired integral by using \texttt{quadv} or \texttt{quadl} with the statements

\begin{verbatim}
p1 = ??; p2 = ??; p3 = ??; ...;
q = quadv( @FunctionName, a, b, [ ], [ ], ...
          p1, p2, p3 )
\end{verbatim}

Assign values to parameters.

Evaluate integral.

Note that the names of the parameters need not be the same at command level as they are in the function M-file. Only the position and data type of the entities must match in the two occurrences, since it is the order and type of items—not the variable names used—that provide the association of values in the two occurrences.

13.8.9 The Off-Axis Electrostatic Potential of Two Rings

Equation (13.36) illustrates a situation in which the integral is a function of an internal parameter. For the integral in Eq. (13.36), for example, we might write the M-file listed in Table 13.6.

We store this file with the name \texttt{rings.m}, then invoke OCTAVE to evaluate and display the desired integral by using \texttt{quadv}, say, with the statement

\begin{verbatim}
q = quadv( 'rings', a, b, [ ], [ ], s )
\end{verbatim}

\(^{25}\)The value zero is interpreted as false; any other value is interpreted as true.

\(^{26}\)Number of function evaluations, left end and width of each subinterval exploited in the process, and the estimated value of the integral over each subinterval.
function y = rings( phi, s )
# RINGS - Integrand for charged rings
# RINGS defines the integrand whose integral gives the
# electrostatic potential in the midplane between two
# uniformly charged circular rings.

    tmp = sqrt( 2.0 ) / ( 2.0 * pi );
    y = tmp ./ sqrt( 2.0 - 2.0*s.*cos(phi) + s*s );
endfunction

Since we seek to explore this integral as a function of $s$ over, say, $0.0 \leq s \leq 4.0$, we must generate a vector of values of $s$ and then, in a loop, invoke quadv once for each element in that vector. We might use the statements

```plaintext
>> s = [ 0.0 : 0.1 : 4.0 ];
>> N = length(s);
>> for i = 1:N
    V(i) = quadv( @rings, 0.0, 2.0*pi, ...
         [], [], s(i) )
>> end

The resulting graph is shown in Fig. 13.12.

13.12 Evaluating Integrals Numerically with Mathematica

Mathematica offers at least two routes for evaluating integrals numerically. For integrals that can be evaluated analytically, we can simply embed the command Integrate within the command N. In response to the simple statement

```plaintext
N[ Integrate[ f, { x, a, b } ] ]
```

where $f$ is an expression that depends on $x$. Mathematica first evaluates $\int_a^b f \, dx$ symbolically and then determines the numerical value from the symbolic result. This same statement will automatically invoke numerical quadrature if a symbolic result cannot be obtained (and all quantities in the integrand have numerical values). If, however, an integral does not admit a symbolic evaluation (or we wish a numeric evaluation even when a symbolic evaluation exists), then we would instead invoke the command NIntegrate in a statement of the form


13.12. EVALUATING INTEGRALS NUMERICALLY WITH MATHEMATICA

Figure 13.12: The potential in the midplane of a charged ring.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{potential_midplane.png}
\caption{Potential in Midplane}
\end{figure}

\texttt{NIntegrate[ f, \{ x, a, b \} ]}

While the defaults assumed by \texttt{NIntegrate} will frequently be acceptable, this command nonetheless possesses several options that give us the ability to override defaults. Among these options are

- \textbf{WorkingPrecision} (default $\$MachinePrecision, often 16), which specifies the number of digits to be used for internal computations.

- \textbf{AccuracyGoal} (default usually figured as half the value of \texttt{WorkingPrecision}), which specifies the absolute accuracy to be sought by stipulating the number of digits to be provided in the solution. \texttt{AccuracyGoal} should not exceed \texttt{WorkingPrecision}.

- \textbf{PrecisionGoal} (default usually figured as half the value of \texttt{WorkingPrecision}), which specifies the relative accuracy sought in the final result. \texttt{PrecisionGoal} should not exceed \texttt{WorkingPrecision}.

\textit{Note:} With \texttt{AccuracyGoal} set to $a$ and \texttt{PrecisionGoal} set to $p$, \texttt{Mathematica} strives for an absolute accuracy in a value $x$ to be less than $10^{-a} + 10^{-p} |x|$.

- \textbf{MaxRecursion} (default automatic), which specifies the maximum number of recursive refinements to be made in seeking the desired precision. This option provides a means to stop the iterations even if goals on accuracy and precision have not been met.

- \textbf{Method} (default usually an adaptive algorithm chosen by \texttt{Mathematica} to be optimum for the integrand), which specifies the method to be used. Among the allowed values are \texttt{GaussKronrodRule}, \texttt{TrapezoidalRule}, and \texttt{MonteCarloRule}.\footnote{A full listing of allowed values can be found by searching for \texttt{NIntegrateIntegrationRules} in the Wolfram Documentation.}

Remember, too, that the statement \texttt{Options[NIntegrate]} will display all options and their current default values for the indicated command. Fuller detail about these features of \texttt{Mathematica} can be found in the \texttt{Mathematica} manuals and in the on-line Help Browser.
With this feature of *Mathematica*, we can evaluate our standard example in many ways. We might, for example, use the statement

\begin{verbatim}
In[1]:= N[ Integrate[ 2*Exp[-t^2]/Sqrt[Pi], { t, 0, 1 } ] ]
\end{verbatim}
\begin{verbatim}
Out[1]= 0.842701
\end{verbatim}

which will evaluate the integral analytically [obtaining erf(1)], evaluate the error function of argument 1, and finally display the result with the default precision intrinsic to the command \texttt{N}. That precision can, of course, be adjusted with the second argument so that, for example, the statement

\begin{verbatim}
In[2]:= N[ Integrate[ 2*Exp[-t^2]/Sqrt[Pi], { t, 0, 1 } ], 20 ]
\end{verbatim}
\begin{verbatim}
Out[2]= 0.84270079294971486934
\end{verbatim}

will yield a result with 20 digits. Alternatively, we might use the statement

\begin{verbatim}
In[3]:= NIntegrate[ 2*Exp[-t^2]/Sqrt[Pi], {t, 0, 1} ]
\end{verbatim}
\begin{verbatim}
Out[3]= 0.842701
\end{verbatim}

which jumps directly to numerical integration, probably via the (adaptive) Gauss-Kronrod method. To illustrate use of some options, we might invoke the statement

\begin{verbatim}
In[4]:= NIntegrate[ 2*Exp[-t^2]/Sqrt[Pi], {t, 0, 1}, Method -> TrapezoidalRule ]
\end{verbatim}
\begin{verbatim}
Out[4]= 0.842701
\end{verbatim}

Reassuringly, these results are all consistent with one another.

### 13.12.1 Quantum Probability

In the opening paragraphs of this section, we have already evaluated the integral that appears in the determination of the probability that a quantum harmonic oscillator in its ground state will be found outside the classical turning point. According to Eq. (13.41), that probability is given by the expression

\begin{equation}
P(|x| > |x_{\text{turn}}|) = 1.0 - \text{erf}(1.0) = 1.0 - 0.842701 = .157299 \tag{13.108}
\end{equation}

which is (probably) correct to all of the digits shown. We conclude that in a bit over 15% of the measurements, the quantum oscillator will be found in the classically forbidden region!

### 13.12.2 The Error Function

To define the function \texttt{gaussian} and then generate successive numerical values of the function \texttt{erf}(x) defined by Eq. (13.26) as \(x\) varies from 0.0 to 3.0 in steps of 0.1, we might use the statements

\begin{verbatim}
In[1]:= gaussian[t_] := 2*Exp[-t^2]/Sqrt[Pi]
In[2]:= uplim = Range[ 0.0, 3.0, 0.1 ];
In[3]:= For[ j=1, j<32, j++, q[j]=NIntegrate[ gaussian[t], {t, 0, uplim[[j]]} ] ]
\end{verbatim}

The values of the upper limits in the list \texttt{uplim} and the values \(q_i\) created by these statements must then be converted into a \textit{list} of the coordinates of the points with a statement like

\begin{verbatim}
In[4]:= pts = Table[ { uplim[[i]], q[i] }, { i, 1, 31 } ];
\end{verbatim}

before the graph of Fig. 13.13 can be created with the statement
13.12. EVALUATING INTEGRALS NUMERICALLY WITH MATHEMATICA

Figure 13.13: Error function.

In[5]:= ListPlot[ pts, PlotJoined -> True, PlotStyle -> Thickness[0.01],
PlotLabel -> "Error Function", GridLines -> Automatic ]

In this final statement, two new options have been stipulated explicitly. Setting PlotJoined to True
instructs Mathematica to connect consecutive points with straight line segments; setting GridLines to Automatic requests the drawing of full-length gridlines at each major tick mark on both axes.

13.12.3 The Off-Axis Electrostatic Potential of Two Rings

Using Mathematica, we could evaluate and plot the integral in Eq. (13.36) to find the electrostatic
potential at the radial coordinate \(s_a\) in the midplane between two identical, uniformly charged
parallel rings of radius \(a\). The statement

In[1]:= rings[\[Phi]_, s_] := Sqrt[2]/(2*Pi*Sqrt[2-2*s*Cos[\[Phi]]+s^2])

defines the integrand, and the statements

In[2]:= s = Range[0.0, 3.0, 0.1];
In[3]:= For[ j=1, j<32, j++, V[j]=NIntegrate[ rings[\[Phi],s[[j]]],
{ \[Phi], 0, 2*Pi } ] ]
In[4]:= pts = Table[ { s[[i]], V[i] }, { i, 1, 31 } ];

establish a vector of the desired parameters, evaluate the integral as a function of \(s\) for \(0.0 \leq s \leq 3.0\)
in steps of 0.1, and create the appropriate list of lists to facilitate plotting. Finally, the statement

In[5]:= ListPlot[ pts, PlotRange -> { 0.0, 1.0 }, PlotJoined -> True,
PlotStyle -> Thickness[0.01], GridLines -> Automatic,
PlotLabel -> "Midplane Potential" ]

produces the graph shown in Fig. 13.14.
13.15 Exercises

13.15.1 ... using Symbolic Methods

13.1. A particle of mass \( m \) moves non-relativistically in one dimension under the action of a constant force \( f \). Starting with Eq. (13.3) and using symbolic integration, find the position \( x \), velocity \( v \), and momentum \( p \) of this particle as functions of time if \( x(0) = x_0 \) and \( v(0) = v_0 \).

13.2. A particle of mass \( m \) moves non-relativistically in one dimension under the action of a constant force \( f \). Starting with Eqs. (13.5) and (13.6) and using symbolic integration, find the position \( x \), velocity \( v \), and momentum \( p \) of this particle as functions of time if \( x(0) = x_0 \) and \( v(0) = v_0 \).

13.3. A particle of mass \( m \) moves non-relativistically in one dimension under the action of a force given by \( f(x) = -kx \), where \( k \) is a (spring) constant. Starting with Eqs. (13.5) and (13.6) and using symbolic integration, find the position \( x \), velocity \( v \), and momentum \( p \) of this particle as functions of time if \( x(0) = x_0 > 0 \) and \( v(0) = 0 \).

13.4. Suppose an object of mass \( m \) moves non-relativistically in one dimension under the action of the force \( f(t) = f_0 e^{-bt} \), where both \( b \) and \( f_0 \) are positive. Let \( x(0) = x_0 \) and \( v(0) = v_0 \). Use symbolic integration to find \( x(t) \) and \( v(t) \) by evaluating the integrals in Eq. (13.3). Then, find and interpret both the limits of these two results as \( t \to \infty \) and the Taylor expansion of these two results for small \( t \).

13.5. The normalized Lorentz distribution function is given by

\[
p(x) = \frac{1}{\pi} \frac{a/2}{x^2 + (a/2)^2}
\]

Using symbolic integration, (a) verify that \( \int_{-\infty}^{\infty} p(x) \, dx = 1 \), (b) evaluate—as best you can—the average \( \bar{x} \) and variance \( \sigma^2 \), defined by

\[
\bar{x} = \lim_{b \to \infty} \int_{-b}^{b} x \, p(x) \, dx \quad \text{and} \quad \sigma^2 = \lim_{b \to \infty} \int_{-b}^{b} (x - \bar{x})^2 \, p(x) \, dx
\]
13.6. Suppose some cataclysmic event stops the earth dead in its tracks and, responding to the sun’s gravitational attraction, the earth falls into the sun. Using symbolic integration, find the time required for the earth to fall over the middle half of its journey to the sun. Expressed in years, what is the value of this time for the earth-sun system? Hint: Since the gravitational potential is $-GmM/x$, conservation of energy yields

$$\frac{1}{2}m\left(\frac{dx}{dt}\right)^2 - G\frac{mM}{x} = -\frac{GmM}{x_0} \quad \implies \quad \frac{dx}{dt} = -\sqrt{2GM}\frac{1}{x} - \frac{1}{x^3}$$

(The negative square root is taken because $x$, the distance to the sun, is known to be decreasing.) This expression then leads to the value

$$T_{\text{midhalf}} = \frac{1}{\sqrt{2GM}} \int_{x_0/4}^{3x_0/4} \left(\frac{1}{x} - \frac{1}{x_0}\right)^{-1/2} dx$$

*Hint:* The evaluation will be simpler if you begin by recasting the problem in dimensionless terms, expressing lengths in units of $x_0$ and times in units of $\sqrt{x_0^3/(2GM)}$. To interpret the significance of this unit of time, determine the period of a circular orbit of radius $x_0$, which will turn out to be $2\pi\sqrt{x_0^3/GM}$. For the earth around the sun, this latter time is, of course, 1 year. *Optional:* Evaluate the time required for the first half of the journey, which involves a convergent but improper integral.

13.7. According to the quantum theory, the probability that the electron in the ground state of the hydrogen atom will be found between the center of the atom and some radius $r$ is given by

$$P(r) = \frac{4}{\pi a^2} \int_0^r e^{-2\rho/a} r' r^2 \rho' d\rho' = 4 \int_0^r e^{-2\rho} \rho^2 d\rho$$

where $a$ is the Bohr radius and $\rho = r'/a$. Evaluate this integral symbolically. Then plot and comment on a graph of $P(r)$ versus $r/a$.

13.8. Consider a source consisting of two uniformly charged disks, each of radius $a$ and each oriented with its center on the $z$ axis and its plane perpendicular to the $z$ axis. Let one disk have its center at $(0,0,ca)$ and carry a positive charge density $\sigma$ and the other have its center at $(0,0,-ca)$ and carry a negative charge density $-\sigma$. Using a symbolic program, show that the on-axis electrostatic potential established by this source is given by

$$V(z) = \frac{\sigma}{2\epsilon_0} \left[\sqrt{a^2 + (z-ca)^2} - \sqrt{a^2 + (z+ca)^2} + |z+ca| - |z-ca|\right]$$

and then explore this potential as a function of $z/a$ for various values of $c$. *Hint:* First find the on-axis potential of a single disk lying in the $xy$ plane, and then construct the desired potential by superposition.

13.9. In a spherically symmetric charge distribution, the charge density is a function of only the radial coordinate, $\rho(r) = \rho(r)$. Suppose $\rho(r) = 0$ for $r > a$. Find the electrostatic potential at a point on the $z$ axis, $\mathbf{r} = z\hat{k}$, for which $z > a$ by setting up and evaluating the integral

$$V(0,0,z) = \frac{1}{4\pi\epsilon_0} \int \frac{\rho(r') dv'}{|r - r'|}$$

symbolically to show that $V(0,0,z) = Q/4\pi\epsilon_0 z$. ($Q$ is the total charge in the distribution.) This result demonstrates formally that the potential at a point outside a spherically symmetric charge distribution can be evaluated by regarding the charge to be concentrated at the center of the distribution. *Hint:* Write $r' = x'\hat{i} + y'\hat{j} + z'\hat{k}$ but then use spherical coordinates to express $x'$, $y'$, and $z'$.  

for this distribution, and (c) find the probability that a single, randomly selected value will lie in the range $-a \leq x \leq a$. Finally, (d) show analytically that you should have expected the result of part (c) to be independent of $a$. *Hint:* Introduce the dimensionless variable $\lambda = x/a$. 

13.9.
13.10. Consider a surface in the $xy$ plane having uniform mass density $\sigma$ and having the shape of a cardioid given in polar coordinates by the function $r(\phi) = a(1 - \cos \phi)$. Using symbolic integration, find (a) the center of mass of this object, (b) the moment of inertia tensor of this object about the $x$, $y$, and $z$ axes, and (c) the radius of gyration about the $z$ axis. *Hints:* The center of mass is defined in Section 13.1.2; the moment of inertia tensor is a $3 \times 3$ tensor whose $ij$ element is given by

$$I_{ij} = \int [(x_1^2 + x_2^2 + x_3^2) \delta_{ij} - x_i x_j] dm$$

where $x_1$, $x_2$, and $x_3$ symbolize $x$, $y$, and $z$, respectively; $\delta_{ij}$ is the Kronecker delta, which has the value 1 when $i = j$ and the value 0 otherwise; and the radius of gyration is defined in Section 13.1.3.

13.11. In quantum mechanics, the two integrals $x_{mn} = \int_{-\infty}^{\infty} \psi_m^*(x) x \psi_n(x) \, dx$ and $p_{mn} = \int_{-\infty}^{\infty} \psi_m^*(x) \frac{\hbar}{i} \frac{\partial}{\partial x} \psi_n(x) \, dx$ are important in a variety of contexts. For a particle in an infinitely deep potential well that extends over the region $-a \leq x \leq a$,

$$\psi_n(x) = \begin{cases} \frac{1}{\sqrt{a}} \cos \frac{n\pi x}{2a} & n = 1, 3, 5, \ldots \\ \frac{1}{\sqrt{a}} \sin \frac{n\pi x}{2a} & n = 2, 4, 6, \ldots \end{cases}$$

Using symbolic integration, show that, for these wave functions, $x_{mn} = 0$ and $p_{mn} = 0$ when $m$ and $n$ are both even or both odd and that

$$x_{mn} = \frac{16a}{\pi^2} (-1)^{(m+n+1)/2} \frac{mn}{(m^2 - n^2)^2}$$

$$p_{mn} = \frac{2i\hbar}{a} (-1)^{(m+n+1)/2} \frac{mn}{(m^2 - n^2)}$$

otherwise. Note that, for purposes of translating the general integrals above to the circumstances of this exercise, the wave functions should both be regarded as zero outside of the interval $-a \leq x \leq a$.

13.12. The sawtooth wave is defined by

$$f(x) = \frac{x}{l} \quad ; \quad -l \leq x \leq l$$

Use symbolic integration to find the Fourier coefficients $a_n$ and $b_n$ for this function and then generate graphs showing the function given by the truncated series

$$f_{\text{trunc}}(x) = a_0 \frac{x}{2} + \sum_{n=1}^{N} \left( a_n \cos \frac{n\pi x}{l} + b_n \sin \frac{n\pi x}{l} \right)$$

for various values of $N$, including 0, 1, 2, 3, and 10.

13.13. The Legendre polynomials $L_n(x)$, $n = 0, 1, 2, \ldots$, are orthogonal on the interval $-1 \leq x \leq 1$ with weight 1. In particular,

$$\int_{-1}^{1} L_m(x) L_n(x) \, dx = \frac{2}{2n + 1} \delta_{mn}$$

Any function defined over the interval $-1 \leq x \leq 1$ can then be expanded in the Legendre series

$$f(x) = \sum_{n=0}^{\infty} c_n L_n(x)$$

(a) Show by hand that the coefficient $c_n$ in this expansion is given by

$$c_n = \frac{2n + 1}{2} \int_{-1}^{1} f(x) L_n(x) \, dx$$
13.15. EXERCISES

(b) Use symbolic integration to find $c_n$ for $n = 0, 1, 2, 3, 4, 5$ and $6$ in the Legendre expansion for the function

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

(c) Graph the functions defined by the partial sums $\sum_{n=0}^{N} c_n L_n(x)$ for $N = 0, 1, 2, 3, 4, 5, 6$.

*Hint:* Quite possibly the symbolic program you are using has the Legendre polynomials built in somehow, and you should study its manuals to find out how to invoke them. Just in case that isn’t true, the first nine Legendre polynomials are

\[
\begin{align*}
L_0(x) &= 1 \\
L_1(x) &= x \\
L_2(x) &= \frac{1}{2}(3x^2 - 1) \\
L_3(x) &= \frac{1}{2}(5x^3 - 3x) \\
L_4(x) &= \frac{1}{8}(35x^4 - 30x^2 + 3) \\
L_5(x) &= \frac{1}{8}(63x^5 - 70x^3 + 15x) \\
L_6(x) &= \frac{1}{16}(231x^6 - 315x^4 + 105x^2 - 5) \\
L_7(x) &= \frac{1}{16}(429x^7 - 693x^5 + 315x^3 - 35x) \\
L_8(x) &= \frac{1}{128}(6435x^8 - 12012x^6 + 6930x^4 - 1260x^2 + 35) \\
L_9(x) &= \frac{1}{128}(12155x^9 - 25740x^7 + 18018x^5 - 4620x^3 + 315x)
\end{align*}
\]

13.15.2 ... using Numerical Methods

13.14. To deduce Simpson’s rule, we start by supposing three consecutive values $f_1$, $f_2$, and $f_3$ of the integrand, where for simplicity in notation we take the points of evaluation to be $x_1 = x_2 - \Delta x$, $x_2$, and $x_3 = x_2 + \Delta x$. Using a symbol manipulating program to do the algebra and calculus, (a) find the coefficients $A$, $B$, and $C$ needed to make the parabola $Ax^2 + Bx + C$ pass through the three points $(x_i, f_i)$, $i = 1, 2, 3$, (b) integrate that parabola over the interval $x_1 < x < x_3$ to find that

$$\int_{x_1}^{x_3} f(x) \, dx \approx \int_{x_1}^{x_3} (Ax^2 + Bx + C) \, dx = \frac{\Delta x}{3} \left( f_1 + 4f_2 + f_3 \right)$$

(c) show that this result actually gives the correct value for $f(x) = x^3$ and, finally, (d) deduce the (extended) Simpson’s rule of Eq. (13.63). *Note:* Because this exercise relates to numerical algorithms, it has been placed in with other exercises that are numerical. This exercise is symbolic, and you should use a symbol manipulating program for parts (a), (b), and (c); however, you should address part (d) by hand.

13.17. (a) Deduce the three-point Gaussian formula, for which—paralleling Eq. (13.88)—we set

$$\int_{-1}^{1} f(x) \, dx = w_1 f(x_1) + w_2 f(x_2) + w_3 f(x_3)$$

We then choose the six quantities $w_i$, and $x_i$ so that the expression gives an exact result for $f(x) = 1, x, x^2, x^3, x^4$, and $x^5$. (b) Verify that the interpolation points $x_i$ and weights $w_i$ are given by Eq. (13.101) where $L_3(x) = (5x^3 - 3x)/2$.

13.18. The (normalized) wave functions for a quantum harmonic oscillator in its first and second excited states ($n = 1$, $n = 2$) are

$$\psi_1(x) = \sqrt{\frac{m \omega}{\pi \hbar}}^{1/4} y e^{-y^2/2}; \quad \psi_2(x) = \frac{1}{\sqrt{2}} \left( \frac{m \omega}{\pi \hbar} \right)^{1/4} (2y^2 - 1) e^{-y^2/2}$$

where $y = x/\sqrt{\hbar \omega/k}$, the energies of these states are $3\hbar \omega/2$, and $5\hbar \omega/2$, respectively, and the symbols have the same meanings as in Section 13.1.8. Find the probability that a harmonic oscillator in each of these states will be found outside the classical turning point.

13.19. The Maxwell-Boltzmann speed distribution yields the integral

$$f(v) = 4\pi \left( \frac{m}{2\pi kT} \right)^{3/2} \int_0^v e^{-mv^2/2kT} v^2 \, dv$$

for the fraction of the molecules having speed less than $v$. Using numerical means, explore this integral as a function of $v$. *Hint:* Re-express the integral using $\sqrt{2kT/m}$ as the unit of velocity.
13.20. Planck’s black body radiation law gives the expression
\[ I(\nu_2, \nu_1) = \frac{8\pi\hbar}{c^3} \int_{\nu_1}^{\nu_2} \frac{\nu^3}{e^{\hbar\nu/kT} - 1} d\nu \]
for the power radiated per unit area in the frequency range \( \nu_1 \leq \nu \leq \nu_2 \). Using numerical means, explore the power radiated in the visible spectrum \( 4 \times 10^{14} \text{ Hz} \leq \nu \leq 7 \times 10^{14} \text{ Hz} \) as a function of temperature. \textit{Hint:} One way to approach this exercise would be to choose a reference frequency \( \nu_0 \) arbitrarily (say \( 10^{14} \text{ Hz} \)) and recast the integral on the dimensionless variable \( s = \nu/\nu_0 \). Examination of \( I \) in units of \( \frac{8\pi\hbar\nu_0^4}{c^3} \) as a function of \( T \) in units of \( \hbar\nu_0/k \) would then be indicated.

13.21. As used in statistical data analysis, the Gaussian distribution for a variable \( t \) is usually expressed in terms of the standard deviation \( \sigma \), the distribution function being
\[ P(a, b) = \frac{1}{\sqrt{2\pi}\sigma} \int_a^b e^{-t^2/(2\sigma^2)} dt \]
Thus, the probability of finding a value between \( a \) and \( b \) is given by
\[ P(a, b) = \frac{1}{\sqrt{2\pi}\sigma} \int_a^b e^{-t^2/(2\sigma^2)} dt \]
Show analytically that \( P(-\sigma, \sigma) = \text{erf}(\sigma/\sqrt{2}\sigma) \), and then evaluate \( P(-3\sigma, 3\sigma) \), \( P(-2\sigma, 2\sigma) \), and \( P(-\sigma, 2\sigma) \) numerically. The values of these three quantities are 0.6827, 0.9545, and 0.9973, respectively—values that give rise to the designations of 68%, 95%, and 99% confidence intervals in statistical data analysis.

13.22. Suppose some cataclysmic event stops the earth dead in its tracks and, responding to the sun’s gravitational attraction, the earth falls into the sun. Using numerical integration, find the time required for the earth to fall over the middle half of its journey to the sun. Expressed in years, what numerically is the value of this time for the earth-sun system? \textit{Hint:} Since the gravitational potential is \(-GmM/x\), conservation of energy yields
\[ \frac{1}{2}m \left( \frac{dx}{dt} \right)^2 - \frac{GmM}{x} = \frac{GmM}{x_0} \quad \Rightarrow \quad \frac{dx}{dt} = -\sqrt{2GM} \sqrt{\frac{1}{x} - \frac{1}{x_0}} \]
(The negative square root is taken because \( x \), the distance to the sun, is known to be decreasing.) This expression then leads to the value
\[ T_{\text{midhalf}} = \frac{1}{\sqrt{2GM}} \int_{x_0/4}^{3x_0/4} \left( \frac{1}{x} - \frac{1}{x_0} \right)^{-1/2} dx \]
\textit{Hint:} The evaluation will be simpler if you begin by recasting the problem in dimensionless terms, expressing lengths in units of \( x_0 \) and times in units of \( \sqrt{x_0^3/(2GM)} \). To interpret the significance of this unit of time, determine the period of a circular orbit of radius \( x_0 \), which will turn out to be \( 2\pi \sqrt{x_0^3/(2GM)} \). For the earth around the sun, this latter time is, of course, 1 year. \textit{Optional:} See if you can develop a means to determine the time required for the first half of the journey, which unfortunately—for numerical approaches—involves a convergent but improper integral.

13.23. The normalized Lorentz distribution function is given by
\[ p(x) = \frac{1}{\pi} \frac{a/2}{x^2 + (a/2)^2} \]
Find the probability that a single, randomly selected value will be in the range \(-a \leq x \leq a\). Make sure to assess the precision of your result by methods that do \textit{not} exploit \textit{a priori} knowledge of the exact value. \textit{Hint:} Before evaluating the integral, introduce the dimensionless variable \( s = x/a \) and note that the result actually doesn’t depend on \( a \), so there is but \textit{one} number to determine.
13.24. According to the quantum theory, the probability that the electron in the ground state of the hydrogen atom will be found between the center of the atom and some radius \( r \) is given by

\[
P(r) = \frac{4}{a^3} \int_0^r e^{-2r/a} r'^2 dr' = 4 \int_0^{r/a} e^{-2\rho^2} d\rho
\]

where \( a \) is the Bohr radius and \( \rho = r/a \). Using numerical integration, evaluate this integral as a function of its upper limit. Then plot and comment on a graph of \( P(r) \) versus \( r/a \).

13.25. The complete elliptic integrals of the first and second kinds are given by

\[
K(k) = \int_0^{\pi/2} \frac{d\phi}{(1-k^2 \sin^2 \phi)^{1/2}} \quad ; \quad E(k) = \int_0^{\pi/2} (1-k^2 \sin^2 \phi)^{1/2} d\phi
\]

Explore these integrals as functions of the modulus \( k \). As part of your exploration, obtain a graph of the period \( T \) of a simple pendulum as a function of the amplitude \( \alpha \) of that pendulum. Analytically, the period of that pendulum is given as a function of \( \alpha \) by \( T/T_0 = (2/\pi)K(\sin(\alpha/2)) \), where \( T_0 \) is the period of the pendulum at small amplitude.

13.26. The angular position \( \theta(t) \) of a simple pendulum swinging with amplitude \( \alpha \) is given by the integral

\[
\omega_0 t = \int_0^\beta \frac{d\phi}{(1-k^2 \sin^2 \phi)^{1/2}}
\]

where, with \( l \) the length of the pendulum and \( g \) the acceleration of gravity, \( \omega_0 = \sqrt{g/l} \), \( k = \sin(\alpha/2) \), and \( \beta = \sin^{-1}((\sin(\theta/2))/k) \). Remember that, because of the choice of signs (see Section 13.1.4), this integral is valid only during the portion of the swing from \( \theta = 0 \) to \( \theta = \alpha \). Obtain graphs of \( \theta \) versus \( \omega_0 t \) over the first quarter of the pendulum’s swing for several different values of \( \alpha \).

13.27. The \( n \)-th order Bessel function can be defined by the integral

\[
J_n(x) = \frac{1}{\pi} \int_0^\pi \cos(n\theta - x \sin \theta) \, d\theta
\]

By evaluating this integral numerically as a function of \( x \) for different values of \( n \), obtain graphs of \( J_0(x) \), \( J_1(x) \), and \( J_2(x) \) over the range \( 0 \leq x \leq 10 \).

13.28. The Bessel function \( J_1(x) \) can be defined by the integral

\[
\frac{1}{x} J_1(x) = \frac{2}{\pi} \int_0^1 \left(1 - u^2\right)^{1/2} \cos(xu) \, du
\]

Using this definition, obtain a graph of \( J_1(x) \) versus \( x \) over the range \( 0 \leq x \leq 10 \).

13.29. A circular ring of radius \( a \) resides in the \( xy \) plane with its center at the origin and carries a charge \( Q \) uniformly distributed about its perimeter. The electrostatic potential established by this ring at an observation point whose cylindrical coordinates are \( (r, \phi, z) \) is

\[
\frac{V(r, \phi, z)}{Q/4\pi \varepsilon_0 a} = \frac{1}{\pi} \int_0^\pi \left(1 - 2 \frac{r}{a} \cos \phi' + \frac{r^2}{a^2} + \frac{z^2}{a^2}\right)^{-1/2} \, d\phi'
\]

Explore this integral as a function of \( r/a \) for several values of \( z/a \).

13.30. A circular current loop of radius \( a \) lies in the \( xy \)-plane with its center at the origin and carries a current \( I' \) counterclockwise as viewed from a point on the positive \( z \) axis. The magnetic field at a point in the \( xz \) plane is given by

\[
\frac{\mathbf{B}(x, z)}{\mu_0 I'/2\pi a} = a^2 \int_0^{\pi/2} \frac{z \cos \phi' \mathbf{i} + (a - x \cos \phi') \mathbf{k}}{|x|^2 + z^2 + a^2 - 2ax \cos \phi' |^{3/2}} \, d\phi'
\]

Explore both components of this magnetic field numerically as functions of \( x/a \) for various values of \( z/a \), including \( z/a = 0.0 \) (which will require some creativity for dealing with the point \( x/a = 1.0 \), at which the integrand diverges at one point in the range of the integration variable).
13.31. In a dimensionless presentation, the intensity in the Fresnel diffraction pattern produced by a single slit when that slit is illuminated by a line source parallel to the slit is proportional to the quantity

\[ I \propto \left| \int_{t_b}^{t_b+\delta t} e^{i\pi t^2/2} dt \right|^2 \]

where \( t \) is measured in a unit determined by the distance of the source from the screen containing the slit, the distance of the observation point from that same screen, and the wavelength of the illuminating radiation. In these units, \( t_b \) locates the position of the lower edge of the slit (or, equivalently, the observation point in the diffraction pattern) and \( \delta t \) measures the width of the slit. Obtain graphs of \( I \) versus \( t_b \)—i.e., graphs of \( I \) versus position on the viewing screen—for various values of \( \delta t \). **Hints:** (1) As a start, let \( t_b \) range over the interval \(-4 \leq t_b \leq 4\) and examine values of \( \delta t \) on the order of 1, but allow these initial explorations to suggest possibly more appropriate values. (2) Note that the real and imaginary parts of the integral appearing in this exercise are related to the integrals defining the Cornu spiral discussed in Section 13.1.6.
Chapter 14

Finding Roots

In this chapter, we seek the roots of a known function \( f(x) \), i.e., we seek values of \( x \) satisfying the equation

\[
f(x) = 0
\]  

(14.1)

If \( f(x) \) is simple, we may be able to find a closed form, analytic solution. More often, however, \( f(x) \) is sufficiently complicated that approximate, numerical methods are needed. Further, many functions will have several roots, some of which may be physically meaningless. Thus, we must learn not only how to find the roots but also how to sort the physically meaningful roots from a possibly larger number of mathematically acceptable ones, the “extras” of which are said to be spurious. We begin this chapter by identifying several physical situations, the full addressing of which requires finding one or more roots of some function. Then we illustrate how to use symbolic algebra systems to approach those that can be addressed analytically, describe a few of many available numerical algorithms, and describe ways to find roots using a variety of numerical approaches and computational tools. Briefly at the end, we comment about the more complicated issue of finding roots of sets of simultaneous linear and non-linear equations. Until that point, our discussion will focus on functions of a single variable.

Whatever the function and whatever the approach, the first step in seeking roots should always be to learn as much as possible about the nature of the function and its roots. Further, since numerical methods in particular—most of them iterative—require a starting guess or guesses and will converge more or less rapidly and reliably depending on the quality of those guesses, a priori knowledge of the approximate location of roots is essential. Thus, we should always start by drawing a graph of \( f(x) \) in sufficient detail to reveal the approximate location of the roots of interest. Since the focus of this chapter is not graphing (and graphing has been fully addressed in earlier chapters), we shall carry out this step once in Section 14.1 as we present several sample problems rather than carrying it out repeatedly in later sections.\(^1\)

14.1 Sample Problems

In this section, we identify several physical contexts in which the essential computational problem is to find the roots of some function, and we obtain graphs of the appropriate functions for later reference.

\(^1\)The graphs could, of course, be produced in any number of ways. Except for Fig. 14.2 (which was produced with \( \text{tgif} \)), the graphs in Section 14.1 have all been produced with IDL. Those in each later section have been produced by whatever software is the subject of that section.
14.1.1 Classical Turning Points

Let $V(x)$ be the potential energy under which an object of mass $m$ is moving in one dimension. Turning points in the motion occur at values of $x$ where the total energy $E$ is entirely potential energy (kinetic energy is zero), i.e., when

$$V(x) = E \quad \text{or} \quad V(x) - E = 0 \quad (14.2)$$

Finding physical turning points thus involves finding the mathematical roots of the function $f(x) = V(x) - E$, i.e., finding solutions to Eq. (14.2). If, for example, the potential energy of interest is given by the cubic polynomial

$$V(x) = \frac{x^3}{10000} + \frac{x^2}{200} - \frac{x}{500} - \frac{1}{2} \quad (14.3)$$

the turning points for the motion of a particle with total energy $E = 0$ moving in this potential energy would satisfy $V(x) = 0$. The first step in finding those turning points would therefore be to produce the graph of Fig. 14.1—a task that may require a bit of trial and error before a suitable range for the independent variable has been found. From this graph, we conclude that $V(x)$ has three real roots, one in the vicinity of $x = -50$, a second in the vicinity of $x = -10$, and a third in the vicinity of $x = +10$. With more refined graphs drawn in the vicinity of each of these roots, we could conclude that the three roots are more tightly bound by the limits

$$-50.00 < x_1 < -47.50 ; \quad -12.50 < x_2 < -10.00 ; \quad 7.50 < x_3 < 10.00 \quad (14.4)$$

14.1.2 A Max-Min Problem: Equilibrium Points

From a different perspective, a particle moving in one dimension under the action of the force $F = -dV/dx$ associated with the potential energy $V(x)$ will be in equilibrium at those points $x$ at which the force is zero, i.e., where

$$F = 0 \implies \frac{dV}{dx} = 0 \quad (14.5)$$

or where the graph of $V(x)$ has a horizontal tangent. Further, evaluated at a point of equilibrium, $d^2V/dx^2 > 0$ implies that the equilibrium is stable while $d^2V/dx^2 < 0$ implies that the equilibrium is unstable. Finding points of physical equilibrium therefore involves finding mathematical roots of the function $F(x) = -dV/dx$, and assessing the stability of those equilibria entails examining the sign of $d^2V/dx^2$.

More specifically, for the potential energy given in Eq. (14.3) and graphed in Fig. 14.1, we would find the points of equilibrium by solving the equation

$$\frac{dV}{dx} = \frac{3x^2}{10000} + \frac{x}{100} - \frac{1}{500} = 0 \quad (14.6)$$

By looking at local extrema in the graph of $V(x)$, we infer that this potential energy exhibits two points of equilibrium, one located in the vicinity of $x = -35$ and the other in the vicinity of $x = 0$. A more refined graph leads to the conclusion that these two roots are bounded by

$$-35.0 < x_4 < -30.0 \quad \text{and} \quad -2.5 < x_5 < 2.5 \quad (14.7)$$

2This function is, of course, not particularly realistic as a potential energy. We can, however, provide a physical context for at least a portion of $V(x)$. If we confine our attention to the region around the one minimum it possesses, we can interpret the function as the potential energy of an anharmonic oscillator, for which—if that minimum occurs at $x = 0$—we might write $V(x) = kx^2/2 + ax^4$ while imposing the constraint that $a$, the coefficient of the cubic perturbation from the potential energy of a simple harmonic oscillator, be small. The expression in Eq. (14.2) simply places the minimum at a different value of $x$ and adds a constant to the potential energy. In what follows, we will explore this function over a wider range of values of $x$ than is physically meaningful. The pedagogic advantage of Eq. (14.2) is that it combines many of the important features of potential energy functions with, as we shall see, tractability by a variety of different approaches.


14.1. SAMPLE PROBLEMS

Figure 14.1: Graph of the potential energy of Eq. (14.3).

Figure 14.2: A system of two masses and three springs. The surface on which the two objects rest is assumed to be frictionless.

14.1.3 Natural Frequencies of Oscillating Systems

Suppose we seek the natural frequencies of oscillation for the system shown in Fig. 14.2, which consists of two objects, each having mass \( m \). Let these objects move in one dimension on a horizontal, frictionless surface, let them be connected to one another with a spring having constant \( k' \), and let each be connected to the nearer wall with a spring having constant \( k \). Further, let the position of each be measured from its equilibrium position. Then, Newton’s second law combined with Hooke’s law leads to the equations of motion\(^3\)

\[
\begin{align*}
    m \frac{d^2 x_1}{dt^2} &= -k x_1 + k' (x_2 - x_1) \quad \text{and} \quad m \frac{d^2 x_2}{dt^2} = -k x_2 - k' (x_2 - x_1) \\
\end{align*}
\] (14.8)

To cast these equations in dimensionless form, we choose a unit of length \( a \), set \( x_i/a = \xi_i \), introduce \( \omega_0 = \sqrt{k/m} \), set \( \kappa = k'/k \), and introduce \( \bar{t} = \omega_0 t \) to find that

\[
\begin{align*}
    \frac{d^2 \xi_1}{d\bar{t}^2} &= -(1 + \kappa) \xi_1 + \kappa \xi_2 \quad \text{and} \quad \frac{d^2 \xi_2}{d\bar{t}^2} = \kappa \xi_1 - (1 + \kappa) \xi_2 \end{align*}
\] (14.9)

Next, seeking sinusoidal (or simple harmonic) solutions, we suppose that

\[
\begin{align*}
    \xi_1(\bar{t}) &= \xi_{10} \cos \omega \bar{t} \quad \text{and} \quad \xi_2(\bar{t}) = \xi_{20} \cos \omega \bar{t} \end{align*}
\] (14.10)

\(^3\)These equations were also discussed in Section 11.1.6.
where the (yet to be determined) frequency $\omega$ is measured in units of $\omega_0$. Substituting these suppositions into Eq. (14.9), we conclude that the (presently unknown) amplitudes must satisfy

$$
\begin{vmatrix}
1 + \pi - \omega^2 & -\pi \\
-\pi & 1 + \pi - \omega^2
\end{vmatrix}
\begin{pmatrix}
\pi_{10} \\
\pi_{20}
\end{pmatrix} = 0
\tag{14.11}
$$

The solution of this equation for the unknowns $\pi_{10}$ and $\pi_{20}$ will be trivial (i.e., both zero—a particularly uninteresting motion) unless the determinant of the matrix of coefficients happens itself to be zero, i.e., unless the characteristic equation

$$
D(\omega) = (1 + \pi - \omega^2 - \pi^2) = \omega^4 - 2\omega^2(1 + \pi) + 1 + 2\pi = 0
\tag{14.12}
$$

is satisfied. We arrive at a fourth-order polynomial $D(\omega)$ in $\omega$, the roots of which will give the natural frequencies of oscillation for the simple system under consideration. Since only even powers of $\omega$ appear, however, we can for the sake of a simpler solution regard the polynomial as quadratic in $\omega^2$.

Suppose we seek the dependence of the roots of this polynomial on $\pi = k'/k$, i.e., on the strength of the middle spring compared to that of the two outer ones. Graphs of $D(\omega)$ versus $\omega$ for four different values of $\pi$ are shown in Fig. 14.3. The lower of the two (positive) roots of $D(\omega)$ appears to be $\omega = 1$ regardless of the value of $\pi$, while the upper of the two roots increases steadily as $\pi$ increases.

14.1.4 Range of Projectile in Viscous Medium

For a fourth example, suppose we seek the range of a projectile of mass $m$ fired with the initial speed $v_0$ at angle $\theta$ up from the horizontal in a viscous medium having damping constant $b$. Supposing the
projectile to move in the \( xy \) plane, with \( x \) horizontal and \( y \) vertical, we begin by invoking Newton’s second law\(^4\) to write the equations of motion

\[
m \frac{d^2x}{dt^2} = -b \frac{dx}{dt} \quad \text{and} \quad m \frac{d^2y}{dt^2} = -mg - b \frac{dy}{dt}
\]

(14.13)

Here, \( g \) is the acceleration of gravity (which we take to be a positive number). These equations are to be solved subject to the initial values

\[
x(0) = 0 \quad , \quad v_x(0) = v_0 \cos \theta \quad , \quad y(0) = 0 \quad \text{and} \quad v_y(0) = v_0 \sin \theta
\]

(14.14)

The solution to this problem—see Chapter 11 on ordinary differential equations—is

\[
x(t) = \frac{mv_0 \cos \theta}{b} \left( 1 - e^{-bt/m} \right)
\]

(14.15)

\[
y(t) = -\frac{mgt}{b} + \frac{m}{b} \left( v_0 \sin \theta + \frac{mg}{b} \right) \left( 1 - e^{-bt/m} \right)
\]

(14.16)

Finally, to find the range \( R \), we seek the value of \( x \) at that non-zero value of \( t \) for which \( y(t) = 0 \). Thus, we need to find the non-zero solution, say \( t_1 \), of the equation

\[
-\frac{mgt}{b} + \frac{m}{b} \left( v_0 \sin \theta + \frac{mg}{b} \right) \left( 1 - e^{-bt/m} \right) = 0
\]

(14.17)

and then evaluate \( x(t_1) \). Equivalently, if we introduce the dimensionless time \( \tau = bt/m \) and the dimensionless parameter \( \alpha = bv_0/(mg) \), the equation whose root we seek becomes

\[
f(\tau) = \tau - (1 + \alpha \sin \theta) \left( 1 - e^{-\tau} \right) = 0
\]

(14.18)

Once we have found the desired root, say \( \tau_1 \), we then determine the range \( R \) by substituting \( \tau_1 \) into Eq. (14.15), finding that

\[
\frac{R(\theta, \alpha)}{v_0^2/g} = \frac{\cos \theta \left( 1 - e^{-\tau_1} \right)}{\alpha} = \frac{\cos \theta}{\alpha(1 + \alpha \sin \theta)} \tau_1
\]

(14.19)

The mathematical task confronting us involves finding the non-zero root of the function \( f(\tau) \) defined in Eq. (14.18). More specifically, finding the angle \( \theta \) at which the projectile should be fired to achieve maximum range would entail

- Choosing a value of \( \alpha = bv_0/(mg) \).
- Finding the non-zero root of \( f(\tau) \) for several values of \( \theta \) ranging from 0 to \( \pi/2 \).
- Calculating and plotting values of \( R(\theta, \alpha) \).
- Finding the value of \( \theta \) corresponding to the peak in a graph of \( R(\theta, \alpha) \) versus \( \theta \).

Note that \( \alpha \) increases as \( b \) (the damping) increases and as \( v_0 \) increases but decreases as \( m \) increases.

To gain some insight into the nature of the desired roots, we begin by plotting the family of graphs of \( f(\tau) \) versus \( \tau \) for various values of \( \theta \) with fixed \( \alpha \). After some exploration using techniques introduced in earlier chapters, we arrive at the several graphs shown in Fig. 14.4. Every graph exhibits a root at \( \tau = 0 \), corresponding to the moment of launch. Though it is hard to judge, each graph for \( \theta = 0^\circ \) also exhibits a second root at \( \tau = 0 \), since a projectile launched at \( \theta = 0^\circ \) returns to its initial altitude immediately. As the angle of launch is increased in each case, however, the second root moves to larger and larger values of \( \tau \). Our task in subsequent sections will be to find numerical values for that second root for various values of \( \alpha \) and \( \theta \), determine the range for each, and find the particular angle at which the range is greatest for a given \( \alpha \).

\(^4\)Compare Eqs. (11.3) and (11.4).
Figure 14.4: Graphs of \( f(\tau) \) versus \( \tau \). The individual frames in this display correspond to different values of \( \alpha \); the graphs in each frame correspond to different values of \( \theta \), ranging in 15° increments from 0° for the highest graph to 90° for the lowest graph in each frame. The scalings have been chosen to reveal the nature of the roots most clearly and are different in each frame.

As a quick aside, we can reassure ourselves that this approach is appropriate by examining the limits of Eqs. (14.18) and (14.19) as \( b \) becomes small. In that limit, both \( \tau = bt/m \) and \( \alpha = bv_0/(mg) \) also become small. To assess the limit, we need to expand Eq. (14.18) to second order in \( \tau \) and \( \alpha \), finding that

\[
0 = \tau - (1 + \alpha \sin \theta) \left[ 1 - \left( 1 - \tau + \frac{\tau^2}{2} \right) \right] = \tau + (1 + \alpha \sin \theta) \left[ -\tau + \frac{\tau^2}{2} \right]
\]

\[
= -\alpha \tau \sin \theta + \frac{\tau^2}{2} = 0 \quad \implies \quad \tau = 2\alpha \sin \theta \quad (14.20)
\]

Then, substituting this result into Eq. (14.19), we find that

\[
\frac{R(\theta, \alpha \to 0)}{v_0^2/g} = \frac{\cos \theta}{\alpha(1 + \alpha \sin \theta)} (2\alpha \sin \theta) = \frac{2 \sin \theta \cos \theta}{(1 + \alpha \sin \theta) \to 2 \sin \theta \cos \theta} \quad (14.21)
\]

which is in complete agreement with the known result for the range of a projectile in the absence of air resistance.
14.1.5 Energy Levels in a Quantum Well

The quantum mechanical analysis for the energy levels of a particle of mass \( m \) in a one-dimensional quantum well characterized by the potential energy

\[
V(x) = \begin{cases} 
\infty & x < 0 \\
-V_0 & 0 \leq x \leq a \\
0 & a \leq x 
\end{cases}
\]  

leads to the conclusion that the allowed energies \( E \) should satisfy the equation\(^5\)

\[
s \cot s = -\sqrt{c^2 - s^2}
\]

where \( c^2 = 2ma^2V_0/h^2 \) and \( s^2 = c^2(1 - E/V_0) \) or \( E = -V_0(1 - s^2/c^2) \). To find the energies, we must thus solve Eq. (14.23) for acceptable values of \( s \) once the depth of the well conveyed by the (fixed) value of \( c \) has been specified.

A simpler version of this equation emerges if we square it (thereby possibly introducing spurious roots because the squared equation is also consistent with a plus sign on the right hand side) and then invoke the trigonometric identity \( 1 + \cot^2 \theta = \sin^{-2} \theta \) to find that

\[
s^2 \cot^2 s = s^2(\sin^{-2} s - 1) = c^2 - s^2 \quad \implies \quad \sin s = \pm \frac{s}{c}
\]

We can be confident that all solutions of Eq. (14.23) will satisfy Eq. (14.24), but we cannot be sure that all solutions of Eq. (14.24)—a potentially larger set—will satisfy Eq. (14.23). Thus, we must in the end remember to sort from all the roots of Eq. (14.24) only those that \( \textit{also} \) satisfy Eq. (14.23).

While we might (see exercises) be interested in the way the allowed energies change as the depth and width of the well change (i.e., as \( c \) changes), we shall here illustrate the techniques by supposing a particular well, namely one for which \( c = 25 \). Then, we seek solutions specifically to

\[
\sin s = \pm 0.04s \quad \text{or} \quad \sin s \mp 0.04s = 0
\]

We suppose we seek roots of this function in the interval \( s \geq 0 \). The graphs in Fig. 14.5 reveal these solutions in several ways. The lower graph in Fig. 14.5(a) shows the function \( \sin s - 0.04s \) (upper sign in Eq. (14.25)), and \( \textit{its} \) roots lie where \( \textit{this} \) graph crosses the \( s \) axis; there are eight roots in this group. Similarly, the upper graph in Fig. 14.5(a) shows the function \( \sin s + 0.04s \) (lower sign in Eq. (14.25)), and \( \textit{its} \) roots appear where the graph crosses the \( s \) axis; there are nine roots in this group. In Fig. 14.5(b), which shows the functions \( \sin s, +0.04s, \) and \(-0.04s\), the roots appear at the values of \( s \) where one or the other of the sloped straight lines intersects the sine curve. From either of these graphs, we conclude that, for the \textit{upper} sign in Eq. (14.25) the eight roots are bounded—crudely—by

\[
\begin{align*}
\text{root 1u} & \quad -0.5 < s < 0.5 \\
\text{root 2u} & \quad 2.5 < s < 3.5 \\
\text{root 3u} & \quad 6.0 < s < 7.0 \\
\text{root 4u} & \quad 8.5 < s < 9.5 \\
\text{root 5u} & \quad 12.5 < s < 13.5 \\
\text{root 6u} & \quad 14.5 < s < 15.5 \\
\text{root 7u} & \quad 19.0 < s < 20.0 \\
\text{root 8u} & \quad 20.5 < s < 21.5
\end{align*}
\]

and that, for the \textit{lower} sign, the nine roots are bounded—again crudely—by

\[
\begin{align*}
\text{root 1l} & \quad -0.5 < s < 0.5 \\
\text{root 2l} & \quad 3.0 < s < 4.0 \\
\text{root 3l} & \quad 5.5 < s < 6.5 \\
\text{root 4l} & \quad 9.5 < s < 10.5 \\
\text{root 5l} & \quad 11.5 < s < 12.5
\end{align*}
\]

\( ^5 \)Text books in quantum mechanics usually treat the finite-depth well for which \( V(x) = -V_0 \) in \(-a \leq x \leq a\) and \( V(x) = 0 \) outside that interval. Nevertheless, the strategy for deriving the result in Eq. (14.23) is described in almost every intermediate level text on quantum mechanics. See, for example, Section 2.6 in David J. Griffiths, \textit{Introduction to Quantum Mechanics} (Prentice Hall, Inc., Upper Saddle River, NJ, 1995). Actually, the condition we choose for illustration gives results identical to those for the \textit{odd} states in the more conventional well.
14.2 Symbolic Approaches

Symbolic approaches to finding roots always take advantage of specific features of the equation or system of equations for which roots are sought. The roots of at least low order polynomials can be found symbolically. Most of us learned in high school algebra, for example, that the (single) root of a linear polynomial is given by

\[ f(x) = ax + b = 0 \implies x = -\frac{b}{a} \quad (14.28) \]

and that the quadratic formula

\[ f(x) = ax^2 + bx + c = 0 \implies x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \frac{2c}{-b \mp \sqrt{b^2 - 4ac}} \quad (14.29) \]

gives the (two) roots of a quadratic polynomial. Similar—though more complicated—expressions exist for the roots of cubic and quartic polynomials. Roots of a polynomial of fifth or higher order cannot be found symbolically unless the polynomial happens to factor into a product of polynomials, each of which individually is of order no higher than quartic. Further, most learned in high school trigonometry that

\[ f(x) = \sin(kx) = 0 \implies x = \frac{n\pi}{k}, \quad n = 0, \pm 1, \pm 2, \ldots \quad (14.30) \]
Occasionally, we will encounter an equation that looks intractable but that can be converted into a tractable case with an appropriate variable transformation. The roots of the expression

\[ f(x) = \frac{a + x}{b + x^2} - c \]  

for example, can be found symbolically because, so long as \( b + x^2 \neq 0 \), the simple recasting

\[ \frac{a + x}{b + x^2} - c = 0 \implies a + x - c(b + x^2) = -cx^2 + x + a - cb = 0 \]  

converts the problem into one involving a quadratic polynomial. Beyond these few cases, very few roots can be found symbolically.

### 14.5 Finding Roots Symbolically with Mathematica

The Mathematica command for solving a wide variety of equations is \texttt{Solve}. It takes two mandatory arguments, the first giving the equation to be solved and the second specifying the variable for which a solution is desired. Thus, the statement

\[
\text{In}[1]:= \text{Solve}[a \cdot x + b == 0, x] \\
\text{Out}[1]= \{\{x \to -\frac{b}{a}\}\}
\]

returns the solution \( x = -b/a \) in the form of a transformation rule and the statement

\[
\text{In}[2]:= \text{Solve}[a \cdot x^2 + b \cdot x + c == 0, x] \\
\text{Out}[2]= \{\{x \to -\frac{b - \sqrt{b^2 - 4ac}}{2a}\}, \{x \to -\frac{b + \sqrt{b^2 - 4ac}}{2a}\}\}
\]

returns a list containing two solutions, each expressed as a transformation rule. The command \texttt{Solve} actually also knows how to deal with cubic and quartic polynomials, though the result in the most general case is likely to be extremely involved and will be useful only if, once obtained, it is quickly converted to floating point form. The command can solve higher order polynomials only if it recognizes a factorization into polynomials, no one of which is of order higher than quartic.

Sometimes (e.g., for roots of polynomials of order greater than three), \texttt{Solve} may be unable to find explicit solutions. In those cases, \texttt{Solve} returns the root in the form \texttt{Root[...]}, e.g.,

\[
\text{In}[3]:= \text{Solve}[a \cdot x^6 - b \cdot x^5 + c == 0, x] \\
\text{Out}[3]= \{\{x \to \text{Root}[c - b \#1^5 + a \#1^6 & , 1]\}, \{x \to \text{Root}[c - b \#1^5 + a \#1^6 & , 2]\}, \ldots\}
\]

(Here, \#1 is a Mathematica-generated generic variable, \texttt{Root[f,n]} stands for the \( n \)-th root of the function \( f \), and the ampersand identifies what precedes it as a “pure” function.\textsuperscript{6}) Basically, this response constitutes an admission by Mathematica that it can’t solve the specified equation symbolically. Fuller description of the command \texttt{Solve}, of a third (optional) argument that specifies variables to be eliminated, and of a few options that affect the behavior of the command are described in the Mathematica manuals. Further, Mathematica will display the options and their current values in response to the statement \texttt{Options[Solve]}.

\textsuperscript{6}See Section 2.2.5 in \textit{The Mathematica Book} (Fourth Edition).
14.5.1 Classical Turning Points

The turning points for the potential energy in Eq. (14.3) are readily found symbolically with the statements

\[
\begin{align*}
\text{In[4]:=} & \quad V = \frac{x^3}{10000} + \frac{x^2}{200} - \frac{x}{500} - \frac{1}{2}; \\
\text{In[5]:=} & \quad \text{soln} = \text{Solve}[V == 0, x] \\
\text{Out[5]} = & \quad \{\{x \rightarrow \text{mess}\}, \{x \rightarrow \text{mess}\}, \{x \rightarrow \text{mess}\}\}
\end{align*}
\]

where the indicated messes, which—because of the presence of \(i = \sqrt{-1}\) here and there—will appear superficially to be complex, are too long and too unfathomable to reproduce here. We can, however, convert this output directly to a (complex) floating point form with the statement

\[
\begin{align*}
\text{In[6]:=} & \quad \text{soln} // \text{N} \\
\text{Out[6]} = & \quad \{\{x \rightarrow 9.3487 + 0.1\}, \{-48.2683 - 8.88178 \times 10^{-16} i\}, \{-11.0804 + 0.1\}\}
\end{align*}
\]

thereby reducing the messes to quantities that are more easily interpreted. The real parts of the roots here obtained are certainly consistent with the bounds given in Eq. (14.4). Further, the assertion that the imaginary parts of the roots are all zero is not unreasonable, given that, by default, the precision of the command \text{N} is on the order of \(10^{-15}\). Indeed, we can support this understanding of the imaginary part by evaluating the roots to more decimal places with the statements

\[
\begin{align*}
\text{In[7]:=} & \quad \text{N[ soln, 20] } \\
\text{Out[7]} = & \quad \{\{x \rightarrow 9.348704197600522706 + 0. \times 10^{-20} i\}, \\
& \quad \{-48.268267566751634131 + 0. \times 10^{-19} i\}, \\
& \quad \{-11.080436630848841393 + 0. \times 10^{-20} i\}\}
\end{align*}
\]

Even at this precision, the assertion that the imaginary parts are all zero is not unreasonable, given that the precision of the command \text{N} is on the order of one in the twentieth digit when its second argument is 20.7

14.5.2 Equilibrium Points

Finding the equilibrium points for the potential energy of Eq. (14.3) is easier. We proceed from the above dialog with the statements

---

7While different machines should give the same results for most of the digits in the real parts of these roots, the values of the imaginary parts may well be quite different from machine to machine, since those values are largely a consequence of internal roundoff.
14.5. FINDING ROOTS SYMBOLICALLY WITH MATHEMATICA

In[8]:= \( dVdx = D[V, x] \);
In[9]:= Solve[ dVdx == 0, x ]
Out[9]= \(
\{ \{ x \to \frac{2}{3} \left( -25 - 8\sqrt{10} \right) \}, \{ x \to \frac{2}{3} \left( -25 + 8\sqrt{10} \right) \} \}\)
In[10]:= soln = N[ % ]
Out[10]= \(
\{ \{ x \to -33.5321 \}, \{ x \to 0.198814 \} \}\)

finding first \( dV/dx \) and then finding the values of \( x \) at which \( dV/dx = 0 \). These results are consistent with the bounds obtained in Eq. (14.7). Finally, to assess the stability of each equilibrium, we evaluate the second derivative at each equilibrium point with the statements

In[11]:= d2Vdx2 = D[ dVdx, x ];
In[12]:= d2Vdx2 /. soln
Out[12]= \{-0.0101193, 0.0101193\}
In[13]:= Quit[]

concluding that the first root \( x = -33.5321 \) locates an unstable equilibrium (\( d^2V/dx^2 < 0; \) local maximum in the potential energy) while the second root \( x = 0.198814 \) locates a stable equilibrium (\( d^2V/dx^2 > 0; \) local minimum in the potential energy).

14.5.3 Natural Frequencies

To find the natural frequencies for the coupled oscillators described in Section 14.1.3, we seek roots of the function \( D(\omega) \) defined in Eq. (14.12). Since \( D(\omega) \) is, in fact, quadratic in \( \omega^2 \), we solve first for \( \omega^2 \)—though we must substitute a single variable \( \lambda \) for \( \omega^2 \) before doing so—and then we take the square root of that solution with the statements

\[
\text{In[1]:=} \quad d = \omega^4 - 2*\omega^2*(1+\kappa) + 1 + 2*\kappa;
\text{In[2]:=} \quad d /. \omega \rightarrow \sqrt{\lambda};
\text{In[3]:=} \quad \text{Solve[} \quad \% \quad = 0, \\lambda \text{]} \]
\text{Out[3]=} \{ \{ \lambda \rightarrow 1 \}, \{ \lambda \rightarrow 1 + 2\kappa \} \}
\text{In[4]:=} \quad \text{freqs =} \quad \sqrt{\lambda} [/\lambda] \quad /\% \]
\text{Out[4]=} \{ 1, \sqrt{1 + 2\kappa} \}

Note—as must on physical grounds be the case—that the solutions obtained for \( \lambda = \omega^2 \) are both positive, so the square roots will be real. (We ignore the negative square roots, since they don’t have the physical significance of the positive square roots.) Note also that the first of the two frequencies is the same for all \( \kappa \) but that the second increases steadily as \( \kappa \)—the strength of the coupling—increases.

To make the point at the end of the last paragraph even more explicitly, we generate a graph of the two frequencies as functions of \( \kappa \) with the statement

---

\(^8\)At this point, for simplicity, we drop the overbar on \( \kappa \).
\(^9\)Evidently, we cannot simply place the list \text{freqs} as the first argument to \text{Plot}. The first argument can be a list, but the list must be constructed within the statement rather than outside of the statement.
Figure 14.6: Natural frequencies for coupled oscillators. The horizontal coordinate is $\kappa = k'/k$ and the vertical coordinate is the frequency in units of $\omega_0 = \sqrt{k/m}$.

14.5.4 Range of Projectile in Viscous Medium

While the problem of finding the range of a projectile in a viscous medium is not effectively addressed by symbolic methods, certain limits of the behavior of that projectile are readily obtained by invoking Mathematica. We begin by defining the trajectory as given by Eqs. (14.15) and (14.16) with the statements

```
In[1]:= x = (m*v0*Cos[\[Theta]]/b) * ( 1 - Exp[-b*t/m] );
In[2]:= y = -m*g*t/b+(m/b)*(v0*Sin[\[Theta]]+m*g/b) * ( 1 - Exp[-b*t/m] );
```

In many instances, the damping constant $b$ is small. Thus, we can sometimes be satisfied with a Taylor expansion of these quantities about $b = 0$, an expansion we obtain with the Mathematica statements

```
To facilitate visualization, some of the output in the following dialog with Mathematica has been recast from the actual form presented by Mathematica.
```

---

10To facilitate visualization, some of the output in the following dialog with Mathematica has been recast from the actual form presented by Mathematica.
14.5. FINDING ROOTS SYMBOLICALLY WITH MATHEMATICA

In[3]:= \text{xb} = \text{Series}[x, \{b, 0, 1\}]

Out[3]= \frac{t^2 v_0 \cos[\theta]}{2m} + O[b]^2

In[4]:= \text{yb} = \text{Simplify[Series[y, \{b, 0, 1\}]]}

Out[4]= -\frac{gt^2}{2} + t v_0 \sin[\theta] + \frac{t^2 (gt - 3v_0 \sin[\theta])b}{6m} + O[b]^2

We notice first—reassuringly—that the terms that are zeroth-order in \textit{b}, i.e., \(x = v_0 t \cos \theta\) and \(y = -\frac{1}{2}gt^2 + v_0 t \sin \theta\), are exactly what we would expect for the projectile in the absence of air resistance. More generally, to find the range when \textit{b} is small, we begin by converting the expression to a polynomial, removing the root at \(t = 0\), and solving for the roots of the Taylor expansion of \textit{y} with the statements

In[5]:= \text{yb} = \text{Normal[yb]};
In[6]:= \text{yb} = \text{Expand[yb/t]};
In[7]:= \text{soln} = \text{Solve[yb == 0, t]}

Out[7]= \{
{t \to \text{mess}},
{t \to \text{mess}}
\}

To sort out which of these two roots is the physically significant one, we evaluate the limit of this expression as \textit{b} \to 0 with the statements

In[8]:= \text{Series[t /. soln[[1]], \{b, 0, 1\}];} \quad \text{Find limit of first solution.}
In[9]:= \text{Simplify[\%, \{m>0, g>0\}]} \quad \text{Find limit of second solution.}

Out[9]= \frac{2v_0 \sin[\theta]}{g} - \frac{2v_0^2 \sin[\theta]^2 b}{3g^2 m} + O[b]^2

Out[11]= \frac{3m}{b} + \frac{v_0 \sin[\theta]}{g} + \frac{2v_0^2 \sin[\theta]^2 b}{3g^2 m} + O[b]^2

Since the second root becomes arbitrarily large as \textit{b} becomes small, we must reject it on physical grounds. Only the first root, which we can display more fully with the statement

In[12]:= \text{sl1} = t /. \text{soln[[1]]};
In[13]:= \text{sl2} = \text{Expand[Simplify[sl1]]}

Out[13]= \frac{3m}{2b} + \frac{3v_0 \sin[\theta]}{2g} - \frac{m \sqrt{9g^2 - 6bv_0 \sin[\theta]}/m + 9b^2 v_0^2 \sin[\theta]^2/m^2}{2bg}

is of interest. Finally, for consistency with approximations already made, we find this root to first order in \textit{b} with the statement

In[14]:= \text{Series[sl2, \{b, 0, 1\}];}
In[15]:= \text{tim = Simplify[\%, \{m>0, g>0\}]} \quad \text{Find limit of second solution.}

Out[14]= \frac{2v_0 \sin[\theta]}{g} - \frac{2v_0^2 \sin[\theta]^2 b}{3g^2 m} + O[b]^2

and find the range by substituting this result into the first order expression for the horizontal coordinate with the statements
In[16]:= rang = Normal[ xb ] /. t -> Normal[tim];
In[17]:= rang = Series[ rang, { b, 0, 1 } ];
In[18]:= rang = Factor[ Normal[ rang ] ]

Out[18] := \( \frac{2v_0^2}{g}\) \( \sin[\theta]\) \( \cos[\theta]\) \left(1 - \frac{4v_0b}{3mg} \sin[\theta]\right)

In[19]:= Quit[]

(We have rearranged this expression a bit from the one presented by Mathematica.) Note in particular that, in the absence of viscous resistance \( (b = 0) \), this result reduces to the known result \( \frac{v_0^2}{g}\) \( \sin[2\theta]\).

14.6 Algorithms for Finding Roots Numerically

In this section we describe several methods for determining a root of a general function numerically.

14.6.1 The Method of Bisection

Suppose, for example, that we know that a single root of \( f(x) \) lies between \( x_{\text{min}} \) and \( x_{\text{max}} \). We can, of course, calculate \( f_{\text{min}} = f(x_{\text{min}}) \) and \( f_{\text{max}} = f(x_{\text{max}}) \). With this input, we can refine our knowledge of the interval in which the root lies by

1. calculating

\[
x_{\text{mid}} = \frac{1}{2}(x_{\text{min}} + x_{\text{max}})
\]

and

\[
f_{\text{mid}} = f(x_{\text{mid}})
\]

2. calculating \( f_{\text{min}} f_{\text{mid}} \), which will be positive if \( f_{\text{min}} \) and \( f_{\text{mid}} \) have the same sign (and the root then lies in the upper half of the original interval) and negative if \( f_{\text{min}} \) and \( f_{\text{mid}} \) have opposite signs (and the root then lies in the lower half of the original interval). Note that we have here assumed the root to be a single root (or at least a root of odd multiplicity), since the criterion invoked depends on the function having opposite signs on opposite sides of the root. (For a root of even order, the function has the same sign on opposite sides of the root, and the criterion here described will not identify the half of the interval in which the root lies. Without embellishment, bisection will fail in this case.)

3. refocusing our attention on the interval \( x_{\text{min}} < x < x_{\text{mid}} \) (i.e., replacing \( x_{\text{max}} \) with \( x_{\text{mid}} \)) or on the interval \( x_{\text{mid}} < x < x_{\text{max}} \) (i.e., replacing \( x_{\text{min}} \) with \( x_{\text{mid}} \)), depending on the outcome of the test at step (2).

4. returning to step (1).

The process is illustrated in Fig. 14.7. In the two cases illustrated, the root lies in the upper half of the original interval, so the second iteration will apply the same procedure to the interval from \( x_{\text{mid}} \) to \( x_{\text{max}} \). With each successive iteration, the interval within which we know the root to lie is shrunk to half of its size at the start of that iteration. This method of bisection is guaranteed to converge provided only that there is a root in the interval, though it will not work if the root is a root of even order and it may be confused if the original interval happens to contain several roots. The iteration is continued until the interval has been reduced to a size that we are willing to accept as a tolerance or, though it happens rarely, if \( f_{\text{mid}} \) actually is zero at some point in the process.
14.6 ALGORITHMS FOR FINDING ROOTS NUMERICALLY

Figure 14.7: The method of bisection. The two parts show the positioning of critical points (a) when the function increases as $x$ increases and (b) when the function decreases as $x$ increases.

Figure 14.8: Newton’s method. The intersection of a tangent line drawn to the curve at the current estimate of the root with the horizontal axis provides the next estimate of the root.

14.6.2 Newton’s Method

A second—more efficient but also less stable—algorithm requires a single starting value rather than a pair of values that bracket a root. Suppose that $x_n$ is the current estimate of the position of the desired root, and let $f_n = f(x_n)$. In Newton’s method, we calculate the next iterate $x_{n+1}$ by

1. calculating the derivative of the function at $x_n$, i.e., $df(x)/dx|_{x=x_n} = f'_n$.
2. extrapolating the tangent line to the graph of the function—a line whose slope is $f'_n$—to the point at which it intersects the horizontal axis, i.e., finding $x_{n+1}$ by requiring it to satisfy

$$\frac{\Delta f}{\Delta x} = \frac{0 - f_n}{x_{n+1} - x_n} = f'_n \implies x_{n+1} = x_n - \frac{f_n}{f'_n} \quad (14.34)$$

The relevant geometry is illustrated in Fig. 14.8. Given an initial “guess”, Newton’s method will converge more rapidly than the method of bisection, though it will have difficulties if the initial guess—or, for that matter, the root itself—is too close to a point at which the derivative of the function is zero. With a poor initial guess, the method may diverge altogether or may converge on a root remote from the one sought.

14.6.3 Other Methods

Numerous other methods for finding roots, some of them restricted to polynomials, have been developed. Because available routines in some programs sometimes use methods other than the two
described above in detail, we include here a brief outline of the main idea in each of several other methods:

- **The secant method** starts from two approximations to the root, say $x_1$ and $x_2$, calculates $f_1 = f(x_1)$ and $f_2 = f(x_2)$, fits a straight line through the points $(x_1, f_1)$ and $(x_2, f_2)$, finds the intersection point $x_3$ of that line with the $x$ axis, and then repeats the process with the two approximations $x_2$ and $x_3$. The method is similar to Newton’s method, except that it uses the values of the function at two points to estimate the slope of the tangent line to the function $f(x)$. It does not require explicit knowledge of the derivative of $f(x)$.

- **Müller’s method** is mildly more sophisticated than the secant method but involves a similar idea. It starts with three estimates of the root, say $x_1$, $x_2$, and $x_3$, calculates $f_1 = f(x_1)$, $f_2 = f(x_2)$, and $f_3 = f(x_3)$, fits a parabola to the points $(x_1, f_1)$, $(x_2, f_2)$, and $(x_3, f_3)$, finds the roots of that parabola using the form of the quadratic formula given second in Eq. (14.29), and finally takes as $x_4$—which replaces $x_1$ for the next iteration—the one root that emerges when the ambiguous sign in Eq. (14.29) is chosen so that the denominator has the larger of the two possible absolute values.

- **Laguerre’s method**, which is limited to polynomials, (1) recognizes that, in terms of its roots $x_i$, a polynomial scaled so that the coefficient of its highest order term is 1 (which has no effect on its roots) can be expressed in the form

$$f(x) = \prod_{i=1}^{n} (x - x_i)$$

(14.35)

(2) finds a couple of relationships satisfied by polynomials in this form, (3) assumes with only weak justification *a priori* that the root sought is isolated from all the others (which are clustered together), and (4) deduces that the next iterate $x_{n+1}$ for the root sought should be determined from the current guess $x_n$ by

$$x_{n+1} = x_n - \frac{N}{G \pm \sqrt{(N-1)(NH-G^2)}}$$

(14.36)

where the sign in the denominator is chosen so that the denominator has the larger of the two possible absolute values, $N$ is the order of the polynomial (equal to the number of roots), and $G$ and $H$ are defined by $G = f'/f$ and $H = (f'/f)^2 - f''/f$, each evaluated at $x = x_n$.

### 14.6.4 Assessing Accuracy

As with all numerical operations, assessing the accuracy of the roots found by an algorithm is essential before we can have confidence in the roots. We can, of course, always substitute the root we have found into the function whose root we seek and simply notice how close to zero the value of the function at the proposed root actually is. In the end, that comparison may provide the most important test of accuracy, though whether the value $f(x_{\text{root}}) = 0.0001$, say, is “close enough” to zero requires a judgment that would take into account the magnitude of $f(x)$ over the important domain of $x$ and, even more, the magnitude of the derivative of $f(x)$ evaluated at the root.\(^{11}\) Certainly, that criterion provides the most accessible test we might apply. It is also a criterion that is easily implemented in a computer program that monitors convergence and stops automatically when the absolute value of the function has been reduced below some specified tolerance.

We would, however, be more interested in the accuracy of the root itself, i.e., in the amount $|x_{\text{exact}} - x_{\text{approx}}|$ by which the (unknown) exact root differs from the root on which a particular

---

\(^{11}\)The less steeply sloped the function at the root, the harder it is to obtain an accurate root by using the value of the function as a guide.
14.9 FINDING ROOTS NUMERICALLY WITH OCTAVE

algorithm converges. Only the method of bisection supports a clear assessment of that difference since, at any particular step in the progress of that algorithm towards a root, we know that the root is “trapped” between two values whose separation is halved with each step in the iteration. Continuing until the difference \(|x_{i+1} - x_i|\) —sometimes called the residual—between consecutive iterates is less than some specified tolerance guarantees that we have located the root to within that tolerance. Implemented in a computer program, an algorithm using the method of bisection can monitor the separation of consecutive iterates and stop when that separation, either as an absolute value or as a fraction of its current value, is reduced below a specified absolute or relative tolerance, respectively.

Given appropriate initial bounds, the method of bisection is guaranteed to converge on a root, but the rate of convergence is slow by comparison with other methods. Assessing the accuracy of the root itself for those other algorithms is more difficult. If, however, the convergence of the algorithm in use is fairly rapid (and Newton’s method will usually satisfy this expectation), taking the (absolute) accuracy of a particular iterate to be on the order of the difference between it and the next iterate, i.e., on the order of the residual, is reasonable. For these algorithms, as with the method of bisection, a computer implementation of a method in the category of this paragraph can also monitor the residual and stop when it has been reduced below a specified (absolute or relative) tolerance. For these methods, however, convergence is not guaranteed, and computer implementations of these methods should include an alternate criterion for stopping the iteration. Such an implementation might, for example, limit the number of iterations and—to keep the user fully informed—display a warning message if the iteration is stopped because this limit is exceeded rather than because the residual has been made sufficiently small. Without some such fail-safe stopping criterion, these methods are in danger of iterating forever.

Beyond monitoring the value of the function or the residual, seeking a particular root by more than one method may give some insight into the accuracy of that root. Different methods have different strengths, weaknesses, and quirks. When they agree to some number of digits, we can have more confidence in the result than we would have if we had obtained it by only one method. When they disagree, we have at least a hint that the function at hand possesses some pathology that we should perhaps attempt to understand before accepting the root we have found.

14.9 Finding Roots Numerically with OCTAVE

Note: All OCTAVE program (.m) files referred to in this chapter are available in the directory \$HEAD/octave, where (as defined in the Local Guide) \$HEAD must be replaced by the appropriate path for your site. At some sites, this directory or some other directory containing these files may also have been placed in OCTAVE’s default search path. If so, the files can be found by OCTAVE without explicit specification of a path. Otherwise, you will have to use the full path to copy them into your default directory to access them.

Functions in OCTAVE for finding the roots of single equations include

- **roots** and **compan**, which find the roots of a polynomial by finding the eigenvalues of an associated companion matrix. The coefficients in the polynomial may be complex. Further, whether the coefficients are complex or real, roots will find both real and complex roots. It is invoked with a statement like

\[
q = \text{roots}(c)
\]

where \(c\) is a (possibly complex) vector containing the coefficients of the terms in the polynomial (highest power first), and \(q\) is a complex vector storing the returned roots. The companion function **compan** calculates the companion matrix associated with the vector of polynomial coefficients. While **compan** can be invoked separately, invocation of **roots** results in an invocation of **compan** behind the scenes. Roots calculates roots to the maximum precision allowed by double-precision arithmetic.
• \texttt{fzero}, which seeks a real root of a (single) nonlinear function—polynomial or otherwise—of a single variable. It is invoked with statements like\footnote{We shall adhere to providing \texttt{FunctionName} by preceding it with an @ sign. OCTAVE will also accept the alternative specification that encloses \texttt{FunctionName} in single or double quotation marks. See footnote 36 in Chapter 4.}

\begin{verbatim}
q = fzero( @func, x )
q = fzero( @func, x, options )
[q, fval, info, output] = fzero(...) 
\end{verbatim}

In these forms,

- \texttt{func} is a string giving the name of a function M-file—built-in or user-defined—that returns the value of the function whose roots are sought.

- \texttt{x} is either a scalar giving a value near the root sought or a two-component vector giving an interval within which a root is to be sought; in either case \texttt{x} must be numeric, not symbolic. If \texttt{x} is a scalar, \texttt{fzero} begins by searching for an interval around \texttt{x} within which \texttt{func} changes sign. Then, taking that interval (or the one specified in a two-component value for \texttt{x}) as a start, \texttt{fzero} refines the interval until the absolute value of the difference between consecutive iterates is no bigger than the default tolerance of $2.2204 \times 10^{-16}$ (which is the smallest difference detectable in double-precision floating point arithmetic).

- \texttt{options}, which is optional, is the name of a variable that has been set to modify the default behavior of \texttt{fzero}. For example, to change the tolerance, we must reset the property \texttt{TolX} by invoking a statement like

\begin{verbatim}
opt = optimset( 'tolx', 1.0e-10 );
\end{verbatim}

to create a variable \texttt{opt} containing the desired value of \texttt{tolx} and then invoke \texttt{fzero} with the expanded statement

\begin{verbatim}
q = fzero( @func, x, opt )
\end{verbatim}

which replaces the default value of \texttt{tolx} with the newly specified value but keeps all other options at their default values.\footnote{\texttt{tolx} is the most commonly used option for \texttt{fzero}. All available options will be displayed by the command \texttt{optimset()} with no arguments. Details on those options are available in the OCTAVE manuals.}

- \texttt{q} stores the returned root.

- \texttt{fval} stores the value of the function at the returned root.

- \texttt{info} stores a returned code with 1 (convergence achieved) and 0 (maximum number of iterations or function evaluations reached without convergence) among the several possible values.

- \texttt{output} is a structure containing the number of iterations, the number of function evaluations and two two-component vectors, one giving the final interval on the horizontal axis within which the root was found and the other giving the interval within which the value of the function lies.

Note also that, should the function whose roots are sought contain parameters, their values must be provided through the use of global variables.

In addition, Section 14.C contains a listing of the Lawrence-written file \texttt{lubisect.m}, which defines an OCTAVE function that uses the method of bisection. It is called with a statement of the form

\begin{verbatim}
q = lubisect( @func, xlb, xub, tol, itmax )
\end{verbatim}
14.9. FINDING ROOTS NUMERICALLY WITH OCTAVE

Table 14.1: The OCTAVE function turnpts.m.

```octave
function fct = turnpts(x)
# TURNPTS: returns values for a particular potential energy.
fct = x^3/10000.0 + x^2/200.0 - x/500.0 - 0.5;
endfunction
```

where `func` is a string giving the name of a function M-file that returns the value of the function whose roots are sought; `xlb` and `xub` give the lower and upper bounds on an interval containing the desired root; `tol` stipulates the largest acceptable difference between consecutive iterates; `itmax` specifies the maximum number of refinements before the search will terminate in case convergence as specified by `tol` fails; and `q` stores the returned root. As the routine is entered, it checks to make sure the function has different signs at the specified end points and, if not, returns to the calling program immediately, setting `q` equal to the entered value of `xlb`. In addition, with each invocation, `lubisect` reports the actual number of iterations used and prints a message of non-convergence if that number exceeds `itmax`.

14.9.1 Classical Turning Points

We might, for example, find the turning points for the potential energy in Eq. (14.3) with the OCTAVE statements

```octave
>> A = [ 1.0/10000.0 1.0/200.0 -1.0/500.0 -0.5 ];
>> q = roots( A )
q = -48.2683
    -11.0804
     9.3487
```

All three roots have turned out to be real. Out of curiosity, one might execute the command

```octave
>> compan( A )
ans =
  -50 20 5000
    1 0 0
    0 1 0
```

to find the companion matrix that `roots` computes in the course of solving the given polynomial.

Other methods for finding the required roots all require that the function be defined in a M-file. For this example, we might construct the file listed in Table 14.1 and stored in the default directory with the name `turnpts.m`. To invoke this method, we also need some information about the location of each root in order to start the process. Basing our input on the information extracted from Fig. 14.1, we would find the roots of the defined polynomial with the OCTAVE statements

```octave
>> format long
>> q = fzero( @turnpts, -50.0 )  Find lowest root.
q = -48.2682675667517
>> q = fzero( @turnpts, -10.0 )  Find middle root.
q = -11.0804366308484
>> q = fzero( @turnpts, 10.0 )   Find highest root.
q =  9.34870419760005
```

Choosing the longer output form for the first root, we find that
format long
>> [ q, fval, info, output ] = fzero( @turnpts, -50.0 )
q = -48.2682675667517
fval = -6.10622663543836e-015
info = 1
output = scalar structure containing the fields:
    iterations = 6
    funcCount = 8
    bracketx = -48.2682675667517 -48.2682675667516
    brackety = -6.10622663543836e-015 8.88178419700125e-016

thus learning a bit more about the process that generated the briefer output.

The same function M-file can alternatively be used in an invocation of lubisect with the statement

>> q = lubisect( @turnpts, -50.0, -47.5, 1.0e-10, 50 )
Number of iterations = 35
q = -48.2682675668184
>> q = lubisect( @turnpts, -50.0, -47.5, 1.0e-8, 50 )
Number of iterations = 28
q = -48.2682675588876

With different starting values, the routine lubisect would, of course, converge on the remaining roots.

### 14.9.2 Natural Frequencies

Because the natural frequencies of the coupled oscillator described in Section 14.1.3 are roots of a polynomial [Eq. (14.12)], we invoke roots. We seek the roots as functions of \( \kappa \), which we shall represent by the variable \( \kappa_{\text{vals}} \) in our statements to OCTAVE. To set a definite objective, we suppose that, ultimately, we would like to have a graph of each of the two roots as a function of \( \kappa \) over the interval \( 0.0 \leq \kappa \leq 4.0 \). and we elect to divide that interval into 40 steps of length 0.1 each. The coefficients of the polynomial are, of course, different for each value of \( \kappa \). Let us therefore begin by creating two arrays, the first of which—a column vector with 41 elements—contains the values of \( \kappa \) and the second of which—an array with 3 columns and 41 rows—contains in each row the coefficients of the polynomial (seen as a quadratic polynomial in \( \omega^2 \)) for the corresponding value of \( \kappa \). These ends are achieved with the statements

```octave
>> kvals = transpose(0.0:0.1:4.0);
>> for i=1:41 unit(i)=1.0; end;
>> unit=transpose(unit);
>> coeffs = [ unit -2.0*(1.0+kvals) 1.0+2.0*kvals ];
```

which set the values of \( \kappa \), create a 41-element column vector \( \text{unit} \) of ones, and then create the 3 column \( \times \) 41 row array of coefficients. Then, we find the roots and concatenate them into a single array with the statements
14.9. FINDING ROOTS NUMERICALLY WITH OCTAVE

Figure 14.9: Natural frequencies as a function of coupling strength. Two normal modes exist. For one, the natural frequency is constant; for the other, the natural frequency increases as the coupling strength increases.

\[
\begin{align*}
\kappa & \quad 0 & 1 & 2 & 3 & 4 \\
\omega & \quad 0 & 0.5 & 1 & 1.5 & 2 & 2.5 & 3 \\
\end{align*}
\]

\[
\begin{align*}
\kappa & \quad 0 & 1 & 2 & 3 & 4 \\
\omega & \quad 0 & 0.5 & 1 & 1.5 & 2 & 2.5 & 3 \\
\end{align*}
\]

```matlab
>> q = [];
>> for i = 1:41 q = [ q roots( coeffs(i,:) ) ]; end;
>> q = transpose(q);
```

The result is a 2 column \( \times 41 \) row array of real values, the first column containing the first root for each value of \( \kappa \) and the second column containing the second root. Finally, since the roots we have at the moment are values of \( \omega^2 \), we take the square root to find the values of \( \omega \) and then plot a graph of each of the roots—one contained in the first column of the array \( \text{roots} \) and the other contained in the second column—with the statements

```matlab
>> q = sqrt( q );
>> plot( kvals, q(:,1), 'color', 'black', 'linewidth', 4 );
>> hold on;
>> plot( kvals, q(:,2), 'color', 'black', 'linewidth', 4 );
>> set( gca, 'ylim', [0.0,3.0], 'fontsize', 16);
>> grid on;
>> xlabel( '\kappa', 'fontsize', 16 );
>> ylabel( '\omega', 'fontsize', 16 );
>> hold off
```

The resulting graph is shown in Fig. 14.9.

### 14.9.3 Range of Projectile

As laid out in Section 14.1.4, finding the range of a projectile moving in a viscous medium begins with finding the time at which a projectile launched at some angle \( \theta \) in a medium characterized by...
a (dimensionless) viscous damping coefficient \( \alpha = \frac{b_0}{mg} \) returns to its initial altitude. In other words, we must find the one non-zero root of the function \( f(\tau) \) defined in Eq. (14.18). Central to the entire calculation is the function M-file listed in Table 14.2 and stored in the default directory in a file named \texttt{projectile.m}. For example, the upper left frame in Fig. 14.4 could have been created by setting several preliminary variables with the statements

\begin{verbatim}
>> global alpha theta
>> thetavals = pi * [0.0:6.0] / 12.0;
>> tau=[0.0:0.01:2.0];
\end{verbatim}

and then executing the statements

\begin{verbatim}
>> alpha = 0.1;
>> theta = thetavals(1);
>> plot( tau, projectile(tau, alpha, theta ), 'color', 'black', 'linewidth', 3.0 )
>> set( gca, 'xlim', [0.0,0.25], 'ylim', [-0.005,0.005], 'fontsize', 14 )
>> hold on;
>> for i=2:7
    theta = thetavals(i);
    plot( tau, projectile( tau, alpha, theta ), 'color', 'black', 'linewidth', 3.0 );
end
>> grid on;
\end{verbatim}

The remaining frames in Fig. 14.4 were produced with similar statements.

For this example, we choose OCTAVE’s built-in routine \texttt{fzero}. Starting with a fresh OCTAVE session, choosing \( \alpha = 0.4 \), setting \( \theta = 30^\circ \) (in radians), referring to Fig. 14.4 to support the initial guess \( \tau = 0.5 \), and accepting (for the moment) all defaults, we find the corresponding time of return to the initial altitude by invoking the statements

\begin{verbatim}
>> alpha = 0.4;
>> theta = thetavals(1);
>> fzero( projectile, 0.5 )
\end{verbatim}
14.9. FINDING ROOTS NUMERICALLY WITH OCTAVE

>> global alpha theta
Set α and θ.
>> alpha = 0.4; theta = pi*30.0/180.0;
Set initial guess for root.
>> tau0 = [0.1 1.0];
>> q = fzero(@projectile, tau0 )
Invoke fzero, displaying result.
q = 0.37644
The corresponding range is then given by Eq. (14.19), which we evaluate and display for the specific case of this paragraph with the statements

>> range = cos(theta) * ( 1-exp(-q) ) / alpha
range = 0.67918
This result is expressed in units of $v_0^2/g$.

To find the angle of fire to attain maximum range, we need to repeat the sample calculation of the previous paragraph for values of $\theta$ ranging from $0^\circ$ to $90^\circ$, choosing increments of $1^\circ$ between consecutive values of $\theta$. Continuing in the same session with OCTAVE, we prepare for a calculational loop with the statements

>> thetadeg = 10.0:1.0:90.0;
Set values of $\theta$ in degrees.
>> thetarad = pi*thetadeg/180.0;
Convert $\theta$ to radians.
Then, choosing $\alpha = 0.4$ and noting from the first frame in Fig. 14.4 that all roots lie between $\tau = 0.1$ and $\tau = 1.0$, we continue with the statements

>> alpha = 0.4; tau0 = [0.1 1.0];
Set $\alpha$; set initial guess for root.
>> soln = [];
Find root for each $\theta$.
>> for i = 1:81
    theta = thetarad(i);
    soln(i) = fzero( @projectile, tau0 );
end
>> range4 = cos(thetarad).*(1-exp(-soln))/alpha;
Evaluate $R$ for each root.
>> plot( thetadeg, range4, 'color', 'black', ...
    'linewidth', 3.0 );
>> grid on;
>> set( gca, 'fontsize', 16 );
>> xlabel( '\theta (deg)', 'fontsize', 16 );
>> ylabel( 'R(\theta)', 'fontsize', 16 );
>> hold on
Repeating this process with $(\alpha = 0.1, \tau_0 = [0.025 0.25])$, with $(\alpha = 0.2, \tau_0 = [0.05 0.4])$, and with $(\alpha = 0.8, \tau_0 = [0.2 1.5])$, storing the results in range1, range2, and range3, respectively, and overplotting each on the axes created above, we finally produce Fig. 14.10. As $\alpha$ increases ($b$ increases, $v_0$ increases, $m$ decreases, or some combination), the maximum range (measured in units of $v_0^2/g$) decreases and the angle of fire to achieve that range becomes shallower and shallower.

Even more specifically, we can examine the values in the four variables range to find the maximum range and the approximate firing angle to achieve it in each case. For example, the statements

>> format long
>> max(range1)
for this function (see Fig. 14.4), we cannot specify a range that begins at 0.0, since the function does not have opposite signs at the two ends of that interval.

16Because the roots for the lowest few values of $\theta$ are difficult to find, a bit of exploration was necessary before discovering that a successful evaluation would emerge only if the evaluations were started at $\theta = 10^\circ$. 
Figure 14.10: Range versus $\theta$ for $\alpha = 0.1$ (highest graph), $\alpha = 0.2$, $\alpha = 0.4$, and $\alpha = 0.8$ (lowest graph).

\[
\begin{array}{cccc}
\text{ans} & 0.913792225626476 \\
\text{>> [ range1(34) range1(35) range1(36) ]} \\
\text{ans} & 0.913563415463742 & 0.913792225626476 & 0.912932244294679
\end{array}
\]

reveal the maximum value of the range for $\alpha = 0.1$ and, further, reveal that that maximum range occurs when the firing angle is near $44^\circ$.\(^{17}\) Fitting a parabola to the three points and finding the coordinate of its maximum yields the angle $43.71^\circ$.\(^{18}\) Similar statements applied to the other three values of $\alpha$ yield that the maximum range (in units of $v_0^2/g$) and firing angle are

\[
\begin{align*}
0.91384 \text{ and about } 44^\circ (43.71^\circ) & \text{ for } \alpha = 0.1, \\
0.84133 \text{ and about } 43^\circ (42.53^\circ) & \text{ for } \alpha = 0.2, \\
0.72602 \text{ and about } 40^\circ (40.43^\circ) & \text{ for } \alpha = 0.4, \text{ and} \\
0.56957 \text{ and about } 37^\circ (37.01^\circ) & \text{ for } \alpha = 0.8.
\end{align*}
\]

14.9.4 Energy Levels in Quantum Well

Finally, we turn to addressing the energy levels for the finite-depth quantum well discussed in Section 14.1.5. We seek solutions to the equations in Eq. (14.25), so we begin by creating the function M-files \texttt{qmupper.m} and \texttt{qmlower.m} listed in Table 14.3 and stored in the default directory. Then, taking the limits as given in Eqs. (14.26) and (14.27), we find the roots in each category with the OCTAVE statements

\[
\begin{array}{cccc}
\text{\texttt{upper}(1) = lubisect( @qmupper, -0.5, 0.5, 1.0e-8, 40 );}} \\
\text{\texttt{upper}(2) = lubisect( @qmupper, 2.5, 3.5, 1.0e-8, 40 );}} \\
\text{\texttt{.}}
\end{array}
\]

\(^{17}\)Remember that $\theta = 10^\circ$ was stored in element 1 of \texttt{thetadeg}, so element 35 corresponds to $\theta = 44^\circ$.

\(^{18}\)See an exercise towards the end of the chapter(s) on specific symbol manipulating programs.
Table 14.3: The OCTAVE functions \texttt{qmupper.m} and \texttt{qmlower.m}.

```octave
function fct = qmupper( s )
# QMUPPER: Returns the value of \(\sin(s) - 0.04\times s\)
  fct = sin(s) - 0.04*s;
endfunction

function fct = qmlower( s )
# QMLOWER: Returns the value of \(\sin(s) + 0.04\times s\)
  fct = sin(s) + 0.04*s;
endfunction

>> upper(8) = lubisect( @qmupper, 20.5, 21.5, 1.0e-8, 40 );
>> lower(1) = lubisect( @qmlower, -0.5, 0.5, 1.0e-8, 40 );
>> lower(2) = lubisect( @qmlower, 3.0, 4.0, 1.0e-8, 40 );
>> 
>> lower(9) = lubisect( @qmlower, 23.5, 24.5, 1.0e-8, 40 );

and display them with the statement

```octave
given by the statements

```octave
> upper
upper =
Columns 1 through 7
 0 3.02048 6.54820 9.05419 13.11879 15.06139 19.76109
Column 8
 20.99429

and the statement

```octave
given by the statements

```octave
> lower
lower =
Columns 1 through 7
Columns 8 through 9
 23.17778 23.86449

and display them with the statement

```octave
> upper.*cos(upper)./sin(upper) + sqrt(25.0^2-upper.^2)
an =
Columns 1 through 7
 NaN 0.0000 48.2540 0.0000 42.5630 0.0000 30.6270
Column 8
 0.0000

and
CHAPTER 14. FINDING ROOTS

>> lower.*cos(lower)./sin(lower) + sqrt(25.0^2-lower.^2)
ans =
Columns 1 through 7
   NaN 49.570 0.0000 45.974 0.0000 37.65 0.0000
Columns 8 through 9
    18.739 0.0000

Only those roots of Eq. (14.25) that give zero (within the precision of our determination of the root) on substitution into Eq. (14.23) can be accepted. We therefore must reject as spurious the values upper(3), upper(5), upper(7), lower(2), lower(4), lower(6), and lower(8). We also reject upper(1) and lower(1); they correspond to evaluation of Eq. (14.23) at s = 0 and, because
\[
\lim_{s \to 0} s \cot s = 1 \quad \text{while} \quad -\sqrt{25^2 - s^2} = -25 \quad \text{when} \quad s = 0,
\]
these roots don’t satisfy the original equation. Thus, we can assemble a single vector containing the physically meaningful roots with the OCTAVE statements

>> s = lower(2:9);
>> for i = 1:2:7 s(i)=upper(i+1); end
>> s

Discard first root in lower.
Replace “bad” roots in s with “good” ones in upper; display “good” roots.

Finally, remembering that we are interested not so much in the values of s but in the associated energies, we exploit the equation
\[
E/V_0 = -(1-s^2/c^2)
\]
to find the allowed energies with the statements

>> energy = -( 1 - s.^2./25.0^2 )
energy =
Columns 1 through 7
-0.985403 -0.941645 -0.868835 -0.767180 -0.637047 -0.479105 -0.294784
Column 8
-0.088777

The statements

>> plot( [0.2, 1.0], [energy(1),energy(1)], 'color', 'black', 'linewidth', 2.0 )
>> hold on;
>> for i=2:8 plot( [0.2,1.0], [ energy(i), energy(i) ], ...
    'color', 'black', 'linewidth', 2.0 ); end
>> set( gca, 'visible', 'off', 'ylim', [-1.5,0.5], 'xlim', [-0.25,1.5], 'fontsize', 16);
>> plot( [0.0,0.1,2.1,2.1.5], [-1.0,-1.0,0.0,0.0], 'color', 'black' );
>> plot( [0.0,0.0,0.0], [-1.5,0.5], 'color', 'black' );
>> text( 0.05, 0.4, 'E/V_0', 'fontsize', 16 );
>> plot( [0.0,0.1], [0.0,0.0], 'color', 'black' );
>> text( -0.15, -1.0, '-1.0', 'fontsize', 16 );
>> text( -0.15, 0.0 , ' 0.0', 'fontsize', 16 );
>> hold off

will produce the energy level diagram of Fig. 14.11. The second of these statements opens the plot window and draws the first energy level, the third freezes the plot window, the fourth adds the remaining levels, and the fifth adjusts the scaling. The remaining statements add labels and an outline of the well.

14.13 Finding Roots Numerically with Mathematica

When Mathematica is able to find an analytic solution for the roots of an equation of interest, numerical values of the resulting solution can be found simply by embedding the command Solve in the command N in a statement of the form

\[
\text{Solve}[\text{equation}, \text{variable}] \rightarrow N[\text{Solve}[\text{equation}, \text{variable}]]
\]
14.13. FINDING ROOTS NUMERICALLY WITH MATHEMATICA

Figure 14.11: Energy level diagram for the one-dimensional quantum well when \( c = 25.0 \). The heavy lines show the energies, measured in units of \( V_0 \); the light lines show the energies of the bottom (−1) and the top (0) of the well.

Here, \( Eqn \) stands for the equation to be solved and \( Var \) represents the variable to be found. If, however, Mathematica is unable to find an analytic solution (or we want to go directly to a numerical approach anyway), we must resort to numerical methods. Mathematica’s workhorse when the function whose roots are sought is a polynomial is the command \( \text{NSolve} \). The syntax for \( \text{NSolve} \) is essentially identical to that of the command \( \text{Solve} \). In its simplest form, we merely invoke the function in a statement like

\[
\text{NSolve}[\text{Eqn}, \text{Var}]
\]

though we might also exploit a further argument in a statement of the form

\[
\text{NSolve}[\text{Eqn}, \text{Var}, \text{NumberOfDigits}]
\]

to stipulate a precision different from the default. More specifically, the statement

\[
\text{In}[1]:= \text{NSolve}[3 x^3 - 2 x + x - 4 == 0, x]
\]

\[
\text{Out}[1]= \{\{x \to -0.600668 - 0.86549 i\},\{x \to -0.600668 + 0.86549 i\},\{x \to 1.20134\}\}
\]

returns numerical values for the three roots of the specified cubic polynomial, including both real and complex roots. Despite its nominal restriction to polynomials, however, \( \text{NSolve} \) is able to find roots for some non-polynomial functions. For example, the statement
CHAPTER 14. FINDING ROOTS

In[2]:= \text{NSolve}\left[\cos(x) == 0, x\right]
\{x \rightarrow \text{ConditionalExpression}[1.(1.5708 + 6.28319 C[1]), C[1] \in \text{Integers}]\}\}\}

reveals that \textit{Mathematica} recognizes the equation \(\cos(x) = 0\) has many roots and successfully if obscurely reports them all. To obtain any particular roots, we would execute statements like

\begin{align*}
\text{In}[3]:= \text{Out}[2] /\ C[1] & \rightarrow 0 \\
\text{Out}[3]= \{\{x \rightarrow -1.5708\},\{x \rightarrow 1.5708\}\} \\
\text{In}[4]:= \text{Out}[2] /\ C[1] & \rightarrow 3 \\
\text{Out}[4]= \{\{x \rightarrow 17.2788\},\{x \rightarrow 20.4204\}\} \\
\end{align*}

\(C[1]\) can, of course, be given either positive or negative integer values. Apparently, when \text{NSolve} is presented with a non-polynomial, it reverts to \text{Solve} followed by \text{N}, and the effort will be successful only if the given equation admits an analytic solution.\textsuperscript{19}

When the function is not a polynomial, when it does not possess an analytic solution, or when we seek a numerically determined solution anyway, we must use the command \text{FindRoot}, which has arguments specifying not only the equation to be solved and the variable sought but also one or two starting values for an iterative search. The general syntax is

\begin{align*}
\text{FindRoot}\left[\text{Eqn},\{\text{Var},\text{Start}1\}\right] \\
\text{FindRoot}\left[\text{Eqn},\{\text{Var},\text{Start}1,\text{Start}2\}\right]
\end{align*}

in which case \text{FindRoot} uses Newton’s method, or

\text{FindRoot}\left[\text{Eqn},\{\text{Var},\text{Start}1,\text{Start}2\}\right]

in which case \text{FindRoot} uses the secant method.\textsuperscript{20} \text{FindRoot} admits several options, the most significant of which are \text{MaxIterations} (default 100) to specify the maximum number of iterations to use before giving up and \text{AccuracyGoal} (default normally half of the value of $\text{WorkingPrecision}$)\textsuperscript{21} to stipulate the accuracy with which the function will be zero at the returned root. Thus, for example, we might seek a particular root of \(\cos(x)\) with the statement

\begin{align*}
\text{In}[3]:= \text{FindRoot}\left[\cos(x) == 0,\{x, 4.0\}\right] \\
\text{Out}[3]= \{\{x \rightarrow 4.71239\}\}
\end{align*}

or the statement

\begin{align*}
\text{In}[4]:= \text{FindRoot}\left[\cos(x) == 0,\{x, 4.0, 5.0\}\right] \\
\text{Out}[4]= \{\{x \rightarrow 4.71239\}\}
\end{align*}

a value we might recognize as \(3\pi/2\). Properly chosen, the additional stipulations in the second argument permit us to select a particular root from the many that we know \textit{a priori} exist.

Out of curiosity, let us test \text{NSolve} with a polynomial that has a multiple root. We execute the statements

\textsuperscript{19}Some earlier versions of \textit{Mathematica} found only two of the many roots of \(\cos(x) = 0\), the command making its own judgment about the interval in which to search and offering no means for us to specify some other interval. For the particular equation \(\cos(x) = 0\), those versions of \textit{Mathematica} also displayed the warning “Solve::ifun: Inverse functions are being used by Solve, so some solutions may not be found."

\textsuperscript{20}Remember that the two starting values given to the secant method are not necessarily lower and upper bounds for the root that will be found. If they do happen to bound a root, however, that is the root that will be found.

\textsuperscript{21}The default for $\text{WorkingPrecision}$ is \$\text{MachinePrecision}$, often 16.
14.13. FINDING ROOTS NUMERICALLY WITH MATHEMATICA

Create polynomial with multiple roots.

Find roots with \texttt{NSolve}.

Clearly, \texttt{NSolve} has returned all four roots of the given polynomial even though one of the three distinct roots is a double root.

14.13.1 Classical Turning Points

To find the turning points for the potential energy in Eq. (14.3), we simply define the polynomial and invoke \texttt{NSolve} with the statements

\[ V = \frac{x^3}{10000} + \frac{x^2}{200} - \frac{x}{500} - \frac{1}{2}; \]

\[ \text{In}[7] := V = x^3/10000 + x^2/200 - x/500 - 1/2; \]
\[ \text{In}[8] := \text{NSolve}[ V == 0, x ]; \]
\[ \text{Out}[8] = \{ \{ x \rightarrow -48.2683 \}, \{ x \rightarrow -11.0804 \}, \{ x \rightarrow 9.3487 \} \}; \]
\[ \text{In}[9] := \text{Quit}[ ]; \]

These results are certainly consistent with the bounds identified in Section 14.1.1.

14.13.2 Range of Projectile

As laid out in Section 14.1.4, finding the range of a projectile moving in a viscous medium begins with finding the time at which a projectile launched at some angle \( \theta \) in a medium characterized by a (dimensionless) viscous damping coefficient \( \alpha = \frac{b v_0}{m g} \) returns to its initial altitude. In other words, we must find the one non-zero root of the function \( f(\tau) \) defined in Eq. (14.18).

To find that root and then the range for a particular choice of \( \alpha \) and \( \theta \), we begin with the statements

\[ \text{In}[1] := \text{rang} = \cos(\Theta) \frac{(1-e^{-\tau}) \cos(\theta)}{\alpha}; \]
\[ \text{Out}[1] = \frac{(1 - e^{-\tau}) \cos(\theta)}{\alpha}; \]
\[ \text{In}[2] := \text{fct} = \tau - (1 + \alpha \sin(\Theta)) (1 - e^{-\tau})(1 + \alpha \sin(\theta)); \]
\[ \text{Out}[2] = \tau - (1 - e^{-\tau})(1 + \alpha \sin(\theta)); \]

which define the range given by Eq. (14.19) and the auxiliary function given by Eq. (14.18). Then, we proceed with the statements

\[ \text{In}[3] := \alpha = 0.4; \ \Theta = \pi*30.0/180.0; \]
\[ \text{In}[4] := \text{FindRoot}[ fct == 0, \{ \tau, 0.2, 0.5 \} ]; \]
\[ \text{Out}[4] = \{ \tau \rightarrow 0.376438 \}; \]
\[ \text{In}[5] := \text{rang} /. \%; \]
\[ \text{Out}[5] = 0.679177 \]

This result is expressed in units of \( v_0^2/g \).

To find the angle of fire to attain maximum range, we need to repeat the sample calculation of the previous paragraph for values of \( \theta \) ranging from 0° to 90°, choosing, say, increments of 1° between consecutive values of \( \theta \). To find the range for each of the selected values of \( \theta \), we invoke the single loop

```mathematica
In[5] := p1 = Expand[ (x-1)^2*(x+3)*(x-4) ]; Create polynomial with multiple roots.
{ 
{ x \rightarrow -3. }, 
{ x \rightarrow 1. }, 
{ x \rightarrow 4. } 
}
Clearly, NSolve has returned all four roots of the given polynomial even though one of the three distinct roots is a double root.

14.13.1 Classical Turning Points

To find the turning points for the potential energy in Eq. (14.3), we simply define the polynomial and invoke NSolve with the statements

In[7] := V = x^3/10000 + x^2/200 - x/500 - 1/2;
In[8] := NSolve[ V == 0, x ];
Out[8] = 
{ 
{ x \rightarrow -48.2683 },
{ x \rightarrow -11.0804 },
{ x \rightarrow 9.3487 } 
};
In[9] := Quit[]

These results are certainly consistent with the bounds identified in Section 14.1.1.

14.13.2 Range of Projectile

As laid out in Section 14.1.4, finding the range of a projectile moving in a viscous medium begins with finding the time at which a projectile launched at some angle \( \theta \) in a medium characterized by a (dimensionless) viscous damping coefficient \( \alpha = \frac{b v_0}{m g} \) returns to its initial altitude. In other words, we must find the one non-zero root of the function \( f(\tau) \) defined in Eq. (14.18). To find that root and then the range for a particular choice of \( \alpha \) and \( \theta \), we begin with the statements

In[1] := rang = \cos(\Theta) \frac{(1-e^{-\tau}) \cos(\theta)}{\alpha};
Out[1] = \frac{(1 - e^{-\tau}) \cos(\theta)}{\alpha};
In[2] := fct = \tau - (1 + \alpha \sin(\Theta)) (1 - e^{-\tau})(1 + \alpha \sin(\theta));
Out[2] = \tau - (1 - e^{-\tau})(1 + \alpha \sin(\theta));

which define the range given by Eq. (14.19) and the auxiliary function given by Eq. (14.18). Then, we proceed with the statements

In[3] := \alpha = 0.4; \ \Theta = \pi*30.0/180.0; Choose \( \alpha \), \( \theta \).
In[4] := FindRoot[ fct == 0, \{ \tau, 0.2, 0.5 \} ]; Find root.
Out[4] = \{ \tau \rightarrow 0.376438 \}
Out[5] = 0.679177
This result is expressed in units of \( v_0^2/g \).

To find the angle of fire to attain maximum range, we need to repeat the sample calculation of the previous paragraph for values of \( \theta \) ranging from 0° to 90°, choosing, say, increments of 1° between consecutive values of \( \theta \). To find the range for each of the selected values of \( \theta \), we invoke the single loop

```
In[6]:= For[ i=0, i<91, i++,
\[Theta] = Pi*i/180.0;
tmp = FindRoot[ fct == 0, {\[Tau], 0.005, 0.8 }];
rng[i+1] = rang /. tmp ]

where we have accepted the default value 100 of MaxIterations. (Remember that \(\alpha\) was set to 0.4 in the previous paragraph.) The starting values are set with reference to Fig. 14.4 to prevent FindRoot from finding the root \(\tau = 0\) (which exists for all values of \(\theta\)). Unfortunately, that strategy doesn’t work for \(\theta = 0\), at which value the only root is \(\tau = 0\). Even though no warning is presented, FindRoot returns an incorrect value for the root at \(\theta = 0\), which we must correct with the statement

In[7]:= rng[1] = 0.0;

after the above loop has terminated. Finally, we produce the graph in Fig. 14.12 with the statements

In[8]:= pts = Table[ {i, rng[i+1]}, {i, 0, 90}];
In[9]:= ListPlot[ pts, PlotJoined -> True,
AxesLabel -> {"theta", "R(theta)"},
PlotStyle -> Thickness[0.005] ]

Further, by noting the three values,

In[10]:= { rng[40], rng[41], rng[42] }
Out[10]= {0.725174, 0.725949, 0.725889 }
In[11]:= Quit[]

we conclude that the maximum range for \(\alpha = 0.4\) occurs near\(^{22}\) \(\theta = 40^\circ\) (or, with parabolic interpolation to find the maximum, \(\theta = 40.43^\circ\)), rather lower than the \(45^\circ\) angle at which that maximum range occurs when \(\alpha = 0\) (no viscous damping).

Had we started with some other value of \(\alpha\), of course, statements identical to those in the previous paragraph would produce a graph corresponding to that other value of the viscous damping.

### 14.13.3 Energy Levels in Quantum Well

Finally, we turn to addressing the energy levels for the finite-depth quantum well discussed in Section 14.1.5, seeking solutions to the equations in Eq. (14.25). In preparation, we first create lists of the lower and upper bounds given in Eqs. (14.26) and (14.27). To that end we execute the statements\(^{23}\)

\[\begin{align*}
\text{In[1]:=} & \text{ ulb = \{-0.5, 2.5, 6.0, 8.5, 12.5, 14.5, 19.0, 20.5 \};} \\
\text{In[2]:=} & \text{ uub = \{0.5, 3.5, 7.0, 9.5, 13.5, 15.5, 20.0, 21.5 \};} \\
\text{In[3]:=} & \text{ llb = \{-0.5, 3.0, 5.5, 9.5, 11.5, 16.0, 17.5, 22.5, 23.5 \};} \\
\text{In[4]:=} & \text{ lub = \{0.5, 4.0, 6.5, 10.5, 12.5, 17.0, 18.5, 23.5, 24.5 \};}
\end{align*}\]

Then, we find the roots in each category with the simple loops

\(^{22}\)Remember that the index of the entries in rng is one larger than the angle in degrees to which that entry corresponds.

\(^{23}\)We use a three-character variable name, where the first character will be u or l for the upper or lower sign in Eq. (14.25) and the remaining characters will be lb or ub for the lower or upper bounds on the root. Thus, for example, ulb will contain the lower bounds for the roots associated with the upper sign.
14.13. FINDING ROOTS NUMERICALLY WITH MATHEMATICA

In[5]:= For[ i = 1, i<9, i++,
 rootu[i] = FindRoot[ Sin[s]-0.04*s, { s, ulb[[i]], uub[[i]]} ] ]
In[6]:= For[ i = 1, i<10, i++,
 rootl[i] = FindRoot[ Sin[s]+0.04*s, { s, llb[[i]], lub[[i]]} ] ]

Next, remembering that the values in rootu[...] and rootl[...] have the form \{s \to \ldots\}, we concatenate the elements into lists and display them with the statements

In[7]:= rtu = s /. Table[ rootu[i], { i, 1, 8 } ]
In[8]:= rtl = s /. Table[ rootl[i], { i, 1, 9 } ]

Remember, however, that physically acceptable solutions must satisfy not Eq. (14.25), which we have solved, but Eq. (14.23). To determine which of these roots are physically meaningful, we substitute each into Eq. (14.23), remembering that \(c = 25\) for the case we have treated. The appropriate statements to Mathematica are\(^{24}\)

\(^{24}\)Remember that values close to zero often should be exactly zero but fail in that regard because of internal roundoff. These values may differ from operating system to operating system.
In[9]:= \texttt{f[x_] := x*Cot[x] + Sqrt[25^2-x^2]}

In[10]:= \texttt{Map[f, rtu]}

\texttt{42.5628, 3.55271 \times 10^{-14}, 30.6267, -4.79616 \times 10^{-14}}\}

In[11]:= \texttt{Map[f, rtl]}

\texttt{42.5628, 3.55271 \times 10^{-14}, 30.6267, -4.79616 \times 10^{-14}}\}

Only those roots of Eq. (14.25) that give zero (within the precision of our determination of the root) on substitution into Eq. (14.23) can be accepted. We therefore must reject as spurious the values \texttt{rtu[[1]], rtu[[3]], rtu[[5]], rtu[[7]], rtl[[1]], rtl[[2]], rtl[[4]], rtl[[6]]}, and \texttt{rtl[[8]]}. Thus, we assemble a single list containing the eight physically meaningful roots with the \texttt{Mathematica} statements

\texttt{In[12]:= s = Rest[ rtl ];}
\texttt{Discard first root in rtl.}
\texttt{In[13]:= For[ i=1, i<8, i=i+2,}
\texttt{s[[i]] = rtu[[i+1]] ]}
\texttt{Replace "bad" roots in s with "good" ones in rtu.}
\texttt{In[14]:= s}
\texttt{Display "good" roots.}


Finally, remembering that we are interested not so much in the values of \texttt{s} as in the associated energies, we exploit the equation $E/V_0 = -(1-s^2/c^2)$ to find the allowed energies with the statement

\texttt{In[15]:= energy = -(1-s^2/25.0^2) \}
\texttt{Out[15]= \{-0.985403, -0.941645, -0.868835, -0.76718, -0.637047, -0.479105, -0.294784, -0.0887774\}}

Then, the statement

\texttt{In[16]:= pts = Table[ \{}0.2, energy[[i]]\}, \{0.8, energy[[i]]\} \}, \{i,8\} \]
\texttt{Out[16]= \{\{}0.2,-0.985403\}, \{0.8,-0.985403\}\}...\}

creates a list, each of whose elements is itself a list of the two points at opposite ends of a single line in an energy level diagram for the system we are studying. Finally, we produce the energy level diagram itself in steps. First, we draw the axes and the lowest energy level with the statement

\texttt{In[17]:= lv[1] = ListPlot[ pts[[1]], PlotJoined -> True,}
\texttt{PlotRange -> \{\{0.0,1.0\},\{-1.5,0.5\}\},}
\texttt{PlotStyle -> Thickness[0.005], AxesOrigin -> \{0.0,-1.5\},}
\texttt{GridLines -> \{\{\},\{-1.0,-0.5,0.0,0.5\},}
\texttt{AxesLabel -> \{"", "E/V0"\} \]}

Then, we create several graphic objects, each containing one of the energy levels, with the statement\textsuperscript{25}

\textsuperscript{25}The list presented as an argument to the command \texttt{Graphics} may contain graphics primitives, e.g., \texttt{Line}, and graphics directives, e.g., \texttt{Thickness}. The order is important, since the directive affects only those primitives that \texttt{follow} it in the list.
14.16 Solving Simultaneous Equations

To this point in this chapter, we have limited ourselves to solving a single equation for a single unknown quantity—though the single equation has frequently exhibited numerous distinct roots. In many contexts, however, the task of finding (one or more) roots will require solving a system of \( n \) simultaneous equations for \( n \) unknowns.\(^{26}\) As with a single equation determining a single unknown, a system of \( n \) equations determining \( n \) unknowns may exhibit more than one solution, each consisting of \( n \) values, one for each of the unknowns. In this section, we merely enumerate a few contexts in which systems of equations arise and outline the strategies for addressing their solution. Fuller discussion can be found in any number of books on linear algebra and/or numerical methods.\(^{27}\)

\(^{26}\)Actually, systems will sometimes be underdetermined (\( n \) equations with \( m \) unknowns, \( n < m \)) or overdetermined (\( n \) equations with \( m \) unknowns, \( n > m \)), but we shall not consider these cases at all.

14.16.1 Systems of Linear Equations

By far the simplest situation occurs when the equations in the system are all linear, i.e., when the unknowns in the system occur only to the first power and never in product with one another. As illustrated in the exercises at the end of this chapter, Kirchoff’s laws in DC circuit theory, least squares fitting of polynomials to experimental data, and some boundary value problems are among the contexts in which linear systems—sometimes quite large linear systems—of equations arise.

Whatever their physical origin, systems of linear equations can most conveniently be presented in matrix form, e.g.,

\[
\begin{pmatrix}
  a_{11} & a_{12} & a_{13} \\
  a_{21} & a_{22} & a_{23} \\
  a_{31} & a_{32} & a_{33}
\end{pmatrix}
\begin{pmatrix}
  x_1 \\
  x_2 \\
  x_3
\end{pmatrix}
= 
\begin{pmatrix}
  b_1 \\
  b_2 \\
  b_3
\end{pmatrix}
\] (14.37)

where, with \( i \) and \( j \) independently assuming the values 1, 2, and 3, the quantities \( x_i \) are the unknowns, the quantities \( b_i \) are the inhomogenieties, and the quantities \( a_{ij} \) are the coefficients defining the equations. Here, we have three equations and three unknowns. More generally, for a system of \( n \) equations and \( n \) unknowns, we would have an \( n \times n \) matrix \( A \) of coefficients, an \( n \) element vector \( x \) of unknowns and an \( n \) element vector \( b \) of inhomogenieties, we would write the equations compactly in the form

\[ Ax = b \] (14.38)

and we would write their solution formally as

\[ x = A^{-1}b \] (14.39)

Less compactly but more usefully (at least occasionally), we might remember Cramer’s rule and write the solution in terms of determinants in the form

\[
x_1 = \frac{| b_1 a_{12} a_{13} |}{ | A | }; \quad x_2 = \frac{| a_{11} b_1 a_{13} |}{ | A | }; \quad x_3 = \frac{| a_{11} a_{12} b_1 |}{ | A | }
\] (14.40)

In this rule, which can be readily extended to \( n \) equations, the denominator in the expression for the \( i \)-th unknown is the determinant of the coefficient matrix and the numerator is the determinant of the matrix created from the coefficient matrix by replacing its \( i \)-th column with the column of inhomogenieties. Cramer’s rule provides a direct, symbolic, and exact solution to the system of \( n \) simultaneous linear equations. Note, however, that the rule gives problems if the determinant of the coefficient matrix happens to be zero or, equivalently, if the inverse \( A^{-1} \) of the coefficient matrix fails to exist. When \( |A| = 0 \), the equations are said to be singular and will either have no solution (equations inhomogeneous; \( b \neq 0 \)) or an infinite number of solutions (equations homogeneous; \( b = 0 \)).

While compact, Cramer’s rule is not particularly useful for numerical solution of even modest sized systems, since the most direct approach to evaluating determinants is vulnerable to roundoff error.28 We can, however, invent alternative methods that are computationally more satisfactory. The simplest algorithm to describe involves Gaussian elimination, in which one variable at a time is systematically eliminated to yield a simpler system whose solution is readily found by a process called backsubstitution. We illustrate with Eq. (14.37), but the schema is readily extended to \( n \) equations. The process of Gaussian elimination entails

---

28 The most direct approach involves sums and differences of products of \( n \) elements taken so that each column and each row is represented once and only once in each product. The result is a combination of terms that are individually large, some of which are positive and some negative. We end up trying to evaluate a difference between two large numbers, an operation that invites roundoff error.
14.16. SOLVING SIMULTANEOUS EQUATIONS

- dividing each equation by the coefficient of $x_1$, obtaining

$$
\begin{pmatrix}
1 & a_{12}' & a_{13}' \\
1 & a_{22}' & a_{23}' \\
1 & a_{32}' & a_{33}'
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
x_3
\end{pmatrix} =
\begin{pmatrix}
b_1' \\
b_2' \\
b_3'
\end{pmatrix}
$$  \hspace{1cm} (14.41)

- keeping the first equation and replacing the second and third with the result of subtracting the first from each in turn, obtaining

$$
\begin{pmatrix}
1 & a_{12}' & a_{13}' \\
0 & a_{22}'' & a_{23}'' \\
0 & a_{32}''' & a_{33}'''
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
x_3
\end{pmatrix} =
\begin{pmatrix}
b_1' \\
b_2'' \\
b_3'''
\end{pmatrix}
$$  \hspace{1cm} (14.42)

- dividing the second and third equations by the coefficient of $x_2$, obtaining

$$
\begin{pmatrix}
1 & a_{12}' & a_{13}' \\
0 & 1 & a_{23}''' \\
0 & 1 & a_{33}'''
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
x_3
\end{pmatrix} =
\begin{pmatrix}
b_1' \\
b_2''' \\
b_3'''
\end{pmatrix}
$$  \hspace{1cm} (14.43)

- keeping the first and second equations but replacing the third with the result of subtracting the second from the third, obtaining

$$
\begin{pmatrix}
1 & a_{12}' & a_{13}' \\
0 & 1 & a_{23}''' \\
0 & 0 & a_{33}'''
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
x_3
\end{pmatrix} =
\begin{pmatrix}
b_1' \\
b_2''' \\
b_3'''
\end{pmatrix}
$$  \hspace{1cm} (14.44)

In essence, Gaussian elimination converts the original system of equations with a general coefficient matrix into an equivalent system whose coefficient matrix is upper triangular. In that form, however, the solution is readily obtained by backsubstitution. The third equation tells us directly that $x_3 = b_3''/a_{33}'$. Then, knowing $x_3$, we find from the second equation that $x_2 = b_2'' - a_{23}'''x_3$ and, knowing $x_3$ and $x_2$, we find from the first equation that $x_1 = b_1' - a_{12}'x_2 - a_{13}'x_3$. The job is done!

Unfortunately, in a computer whose arithmetic is done to finite precision, the order in which the equations are treated in this process and the order in which the variables are placed can have a significant impact on the quality of the solution obtained. Thus, while Gaussian elimination with backsubstitution provides a starting point, it requires sophisticated embellishment to choose the optimum equation and variable to be the focus of each step in the process. Effecting that embellishment entails a process called pivoting, in which at each step we examine the coefficients in the remaining equations and reorder either the equations (partial pivoting) or the equations and the variables (full pivoting) to optimize the accuracy of the solution. Gaussian elimination with pivoting (alternatively called Gaussian elimination with pivotal condensation) yields a more involved program but also increases the likelihood of useful results.

A similar strategy exploits the property that, under appropriate (and not too restrictive) conditions, the coefficient matrix $A$ can be factored into a product of two matrices $L$ and $U$, the first of which has non-zero elements only on and below the main diagonal (and the elements on the main diagonal are all ones) and the second of which has non-zero elements only on and above the main diagonal, and the two matrices are unique.\textsuperscript{30} That is, we can write

$$
A = LU \quad \text{where} \quad L = \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix} \quad \text{and} \quad U = \begin{pmatrix}
u_{11} & u_{12} & u_{13} \\
u_{21} & u_{22} & u_{23} \\
u_{31} & u_{32} & u_{33}
\end{pmatrix}
$$  \hspace{1cm} (14.45)

\textsuperscript{29}For the sake of simplicity, we will not bother to keep track of the relationship between the original coefficients and those generated along the way. The nature of the algorithm will be clear even without that knowledge.

This so-called *LU* decomposition allows us to seek the solution of the original equation in steps. First, we view the equation in the form

$$Ax = b \implies LUx = b \implies Ly = b \text{ where } y = Ux$$  \hspace{2cm} (14.46)

Since the first equation in the (matrix) equation $Ly = b$ tells us $y_1$ directly, the second tells us $y_2$ directly once $y_1$ is known, and the third tells us $y_3$ directly once $y_1$ and $y_2$ are known, the equation $Ly = b$ is readily solved for $y$. Then, however, once $y$ is known, a similar process that starts with $x_3$, then moves to $x_2$ and finally to $x_1$ directly solves the equation $Ux = y$ for the original unknowns in $x$.

The methods outlined in the previous paragraphs will all work for (almost) any system of linear equations. Sometimes, however, the coefficient matrix may have special properties that can be exploited to simplify the algorithm or—important for large systems—reduce the requirements on memory for storage of matrices and intermediate results. The coefficient matrix may be symmetric ($a_{ij} = a_{ji}$) or it may be sparse (only a small fraction of its elements differing from zero). In the latter category, the matrix may be tridiagonal (non-zero elements only on the main diagonal, on the diagonal immediately above the main diagonal, and on the diagonal immediately below the main diagonal). Special savings in storage can be achieved if the matrix happens to be both symmetric and tridiagonal, since in that case the only elements that need be stored are those on the main diagonal and those on the diagonal immediately above the main diagonal, $2n - 1$ elements rather than $n^2$ elements.

The methods outlined in the previous paragraphs are also all *direct* methods, i.e., each leads directly to the desired solution in a finite number of steps. Except for roundoff error, each would yield an exact solution to the equations at hand. When the coefficient matrix is sparse, an *iterative* approach may be computationally more efficient. Such approaches entail finding a means by which an initial guess can be repeatedly refined until some criterion of convergence has been met. Among the most common such procedures applies to Laplace’s equation. As judiciously as possible, we “guess” a solution at a regular grid of points laid over the domain of the problem. Then, to carry out the first iteration, we step systematically through that grid, replacing the value at each grid point with the average of the values at its nearest neighbors. Then, we repeat the process with the results of the first iteration as input, generating the second iteration, and continuing until—say—no value changes by more than some specified tolerance. The only drawback to an iterative method is that we have added the uncertainties associated with convergence to those potentially generated by computer roundoff. In terms of computational labor, however, the sacrifice is often worth the gains.

Finally, we merely mention the more sophisticated approach—*singular value decomposition*—we must adopt if the system of equations confronting us happens to be nearly singular, in which case the presence of computer roundoff introduces instabilities in the simpler methods.\(^{31}\)

Each of the software packages described in this book makes available a spectrum of commands to solve simultaneous linear equations. We here merely enumerate those commands and indicate the general strategy each implements. We leave the descriptive details to the vendor-supplied documentation, and we leave illustrative applications to the exercises at the end of the chapter.

**MATHEMATICA:** Within *Mathematica*, systems of linear equations can be solved analytically either with the special command *LinearSolve*, which accepts the matrix of coefficients and the vector of inhomogeneities as arguments, or with the command *Solve*, whose arguments then are lists of the equations and the unknowns. Numerical solutions can be found by nesting either *LinearSolve* or *Solve* in the command *N*. While *NSolve* cannot handle systems of equations, *FindRoot* can address systems of equations if we simply make its first argument a list of the equations to be solved, supply as its second argument a list containing the first variable and a starting guess for its value in the solution, supply as its third argument a list containing the second variable and a starting guess for its value in the solution, etc.

---

14.16. SOLVING SIMULTANEOUS EQUATIONS

**OCTAVE:** The simplest route in OCTAVE to find a numerical solution to a system of linear equations was illustrated in Section 4.3.4 and invokes either the operator \( \backslash \), which gives the solution to the equation \( Ax = b \) as \( x = A \backslash b \), or the operator /, which give the solution to the equation \( xA = b \) as \( x = b/A \).

### 14.16.2 Systems of Nonlinear Equations

When one or more of the equations in a system to be solved is nonlinear, the task is much more difficult. Occasionally, systematic elimination of one variable at a time, followed by back-substitution, will yield a solution. More often, the system is not amenable to such a simple approach. We must resort to more involved approaches. Fitting experimental data by the method of least squares to functions in which the parameters do not appear linearly and finding the points of equilibrium in a system with more than one independent variable are common sources of such problems.

Since numerical methods for finding the root (or roots) of a nonlinear system of equations are all iterative, possession of a good starting guess is imperative. In two dimensions, where the equations to be solved are

\[
\begin{align*}
 f_1(x_1, x_2) &= 0 \quad ; \quad f_2(x_1, x_2) = 0 
\end{align*}
\]

we might begin by drawing a map in the \((x_1, x_2)\) plane showing the zero contours of each function. The map in Fig. 14.14, for example, shows the zero contours for the two functions

\[
\begin{align*}
 f_1(x_1, x_2) &= \sin \left( \frac{x_1^2}{20} \right) - \cos \left( \frac{x_2}{5} \right) \quad ; \quad f_2(x_1, x_2) = x_1 \tanh(x_2) - 5.0 
\end{align*}
\]

The actual intersections of the dashed and solid curves near the labels A reveal two roots. In addition, the dashed and solid curves pass close to one another—but do not actually intersect—in the vicinity of the labels B. The roots near A can probably be found relatively easily by an iterative search procedure. That there are “almost” roots near B may confuse some algorithms and, if those points are close enough to the real roots near A, they might even cause difficulties in finding the real roots.

As the number of independent variables increases, the search described in the previous paragraph would move from intersection points of curves in a plane to intersection points of three surfaces in three-space to intersection points of four hypersurfaces in four-space to .... Sometimes it may be possible (and wise) to solve for some of the variables in terms of the others and temporarily eliminate some variables (i.e., reduce the dimensionality of the search). The task is complicated and, beyond the simple suggestion of striving to reduce the dimensionality, no general guidelines can be given. Any means, however devious, that can be exploited to give clues as to the existence of roots and, even better, to their whereabouts should be exploited as fully as possible before actually embarking on a numerical search.

Once a root has been located approximately, we might adopt a brute force technique, writing a program that

1. Accepts a guessed solution, one value for each unknown,
2. Evaluates the functions and displays the result, and
3. Returns to step 1 for a new guess.

On first pass, we enter the initial guess. Then, after seeing how well that guess works, we make a second (informed or, maybe, random) guess, repeating the process until the values of all functions have been reduced to acceptably small levels. Depending on the dimensionality of the search, we will usually develop a feel for the effect of changes in each member of the guessed solution. Fairly quickly, we may develop a skill at narrowing in on an acceptable solution.
Figure 14.14: Zero contours for the two functions in Eq. (14.48). The function $f_1(x_1, x_2)$ is shown with solid lines; $f_2(x_1, x_2)$ is shown with dashed lines. The dashed curve actually intersects the solid curve in two points near the labels A; the two curves *almost* intersect near the labels B.

More systematic searches in multi-dimensional parameter spaces are harder to design. One route in particular expands Newton’s method. Suppose, to be specific, we seek solutions to the three nonlinear equations

$$f_1(x_1, x_2, x_3) = 0 \; ; \; f_2(x_1, x_2, x_3) = 0 \; ; \; f_3(x_1, x_2, x_3) = 0$$  \hspace{1cm} (14.49)

Suppose, further, that we have examined the equations and determined that there does indeed exist a root in the immediate vicinity of the point $(x_1^{(0)}, x_2^{(0)}, x_3^{(0)})$. We might then suppose that the actual root differs from this guess by a small amount, say,

$$x_1 = x_1^{(0)} + \delta x_1 \; ; \; x_2 = x_2^{(0)} + \delta x_2 \; ; \; x_3 = x_3^{(0)} + \delta x_3$$  \hspace{1cm} (14.50)

and demand that the “corrections” satisfy the equations

$$0 = f_1(x_1^{(0)} + \delta x_1, x_2^{(0)} + \delta x_2, x_3^{(0)} + \delta x_3)$$

$$0 = f_2(x_1^{(0)} + \delta x_1, x_2^{(0)} + \delta x_2, x_3^{(0)} + \delta x_3)$$

$$0 = f_3(x_1^{(0)} + \delta x_1, x_2^{(0)} + \delta x_2, x_3^{(0)} + \delta x_3)$$  \hspace{1cm} (14.51)

These equations are, of course, not really any more tractable as they stand than were the original equations. Because the corrections are all presumed small, however, we should be able to approximate these equations by expanding each in a three dimensional Taylor series. Keeping only the linear terms, we find—at least approximately—that

$$0 = f_1 + \frac{\partial f_1}{\partial x_1} \delta x_1 + \frac{\partial f_1}{\partial x_2} \delta x_2 + \frac{\partial f_1}{\partial x_3} \delta x_3$$

$$0 = f_2 + \frac{\partial f_2}{\partial x_1} \delta x_1 + \frac{\partial f_2}{\partial x_2} \delta x_2 + \frac{\partial f_2}{\partial x_3} \delta x_3$$  \hspace{1cm} (14.52)
Here, \( f_1, f_2, f_3 \), and all the derivatives are evaluated at \( x^{(0)}_1, x^{(0)}_2, x^{(0)}_3 \) and are known. In effect, we have converted the problem of finding \( \delta x_1, \delta x_2 \) and \( \delta x_3 \) into one of solving a set of simultaneous linear equations! Once that solution is in hand, we take \( x^{(1)}_1 = x^{(0)}_1 + \delta x_1, \ldots \) as a refined approximation to the desired solution and repeat the process to obtain \( x^{(2)}_1, \ldots \), continuing until some chosen convergence criterion is satisfied.

Each of the software packages described in this book makes available a spectrum of commands to solve simultaneous nonlinear equations. We here merely enumerate those commands and indicate the general strategy each implements. We leave the descriptive details to the vendor-supplied documentation, and we leave illustrative applications to the exercises at the end of the chapter.

**MATHEMATICA:** Mathematica’s command `Solve` is capable of solving some sets of nonlinear equations analytically. In that case, the first argument to `Solve` becomes a list containing the equations to be solved and the second argument becomes a list containing the variables for which solutions are sought. When `Solve` can find solutions to a system of nonlinear equations, nesting `Solve` in the command `N` will find numerical solutions. Whether `Solve` is successful or not, numerical solutions to systems of nonlinear equations can be found with `FindRoot`. We simply make its first argument a list of the equations to be solved. Then, for each unknown, we add one further argument in the form of a list containing (first) the name of the variable and (second) an initial guess for the value of that variable in the solution.

**OCTAVE:** For numerical solution of systems of nonlinear equations, OCTAVE makes available the function `fsolve`, which uses an iterative approach. The function must therefore be supplied with an initial guess for the roots.

### 14.17 Exercises

#### 14.17.1 ... using Symbolic Methods

14.1. Derive both forms of the quadratic formula as given in Eq. (14.29) for the roots of the polynomial \( ax^2 + bx + c \). **Hint:** Start by completing the square, i.e., by adding and subtracting the right amount so that the polynomial can be expressed in the form \( a(x - \alpha)^2 + \beta \).

14.2. Find the value of \( x \) at which \( f(x) = ax^2 + bx + c \) has an extremum and determine a criterion involving the coefficients (or some of them) for deciding whether the extremum is a maximum or a minimum. Assume that \( a, b, \) and \( c \) are real.

14.3. Find the points at which the function \( f(x) = ax^3 + bx^2 + cx + d \) has (local) extrema and find a criterion involving the coefficients (or some of them) that will assure that the function has three real roots. Assume that \( a, b, c, \) and \( d \) are real.

14.4. In some quantum calculations, the need to solve the equation \( x(x + 1) = l(l + 1) \) for \( x \) arises. Find those roots, noting particularly that, since the equation is quadratic, there are two roots. The obvious root \( x = l \) is not the only one.

14.5. Each of the three blades of a lawn mower has radius \( a \). As shown in Fig. 14.15, the center blade is invariably mounted somewhat in front of the two outside ones so that the areas cut by each blade can overlap without risking collision of the blades with one another. What must be the minimum offset \( x \) of the center of the middle blade from the line joining the centers of the two outer blades so that their cutting paths will overlap by an amount \( y \) without collision of the blades?

14.6. For the projectile discussed in Section 14.1.4, generate a family of graphs showing \( \tau(\theta) \) as a function of \( \theta \) for selected fixed values of \( \alpha \).
In Section 14.1.4, we deduced that the range of a projectile of mass $m$ launched with speed $v_0$ at an angle $\theta$ up from the horizontal in a medium characterized by a (linear) viscous damping coefficient $b$ could be found by (1) finding the non-zero root of the equation
\[ f(\tau) = \tau - (1 + \alpha \sin \theta)(1 - e^{-\tau}) = 0 \]
where $\alpha =bv_0/(mg)$, and then (2) evaluating the range from the expression
\[ R(\theta, \alpha) = \frac{v_0^2}{g} \frac{\cos \theta(1 - e^{-\tau})}{\alpha(1 + \alpha \sin \theta)} \]
We could view the first of these equations as defining the function $\tau(\theta)$ implicitly. In principle, for a given $\alpha$, we could imagine solving the first equation explicitly for $\tau$ as a function of $\theta$ and substituting that solution into the second equation to find an expression for the range—again for a given $\alpha$—explicitly as a function of $\theta$ alone. If we had that expression in hand, we would find the maximum range by solving the equation $dR(\theta)/d\theta = 0$ for $\theta$ and then evaluate the range at that specific value of $\theta$. In the absence of that expression, we can still differentiate $R(\theta)$ implicitly, recognizing the (hidden) dependence of $\tau$ on $\theta$, and we can differentiate $f(\tau)$. The resulting equations together might then be combinable in a way that would lead to a determination of the maximum range more directly than the route described in the text. Pursue this idea, using a symbolic manipulating program as much as possible to simplify the calculation. The ultimate objective would be to deduce and test a procedure that leads to a numerical value for the maximum range of this projectile when $\alpha$ is given. Note: This exercise is almost certainly difficult and potentially frustrating. The author has no idea whether success awaits the persistent in this endeavor.

14.8. Find the natural frequencies for the three modes of oscillation characterizing the system that results when the system shown in Fig. 14.2 is extended to contain three objects coupled in a line. Take the four springs all to have the same spring constant but allow for the possibility that the middle object may have a mass $m'$ different from the mass $m$ of the two outside objects. In particular, measure frequencies in units of $\sqrt{k/m}$ and seek a graph showing the frequency of each of the modes as a function of $\beta = m'/m$. Hint: To help you get started and to facilitate focusing on the solution of the ODEs rather than on deriving them, note that, for three masses, the equations of motion will be
\[ m \frac{d^2 x_1}{dt^2} = -k x_1 + k(x_2 - x_1) \]
\[ m' \frac{d^2 x_2}{dt^2} = -k(x_2 - x_1) + k(x_3 - x_2) \]
\[ m \frac{d^2 x_3}{dt^2} = -k(x_3 - x_2) - kx_3 \]

14.9. In dimensionless units, the energy shifts $E$ that occur in the states of hydrogen for $n = 3$ when an external constant electric field is turned on are given by the roots of the ninth-order polynomial
\[ f(E) = E^9 - \frac{243}{2} E^7 + \frac{59049}{16} E^5 - \frac{531441}{16} E^3 \]
Find the distinct roots of this polynomial and the multiplicity of each root.

14.10. If, when divided by a single line into a square and a rectangle, the resulting smaller rectangle has the same aspect ratio as the original rectangle, the rectangle is called a golden rectangle and the ratio of the longer to the shorter side is called the golden ratio. Rectangles having that ratio are considered to be particularly aesthetic and can be found in many places, including ancient Greek architecture and, intriguingly, in standard, present-day plastic credit cards. Express the above defining criterion to find a quadratic equation for the golden ratio and then solve the equation to find that ratio.

14.17.2 ... using Numerical Methods

14.11. Use numerical methods to solve Eq. (14.6) for the equilibrium points in the potential energy in Eq. (14.3).

14.12. One way to find the square root of a (positive) number $a$ is to find the root of the function $f(x) = x^2 - a$. (a) Apply Newton’s method symbolically to show that $x_{n+1} = (x_n + a/x_n)/2$. (b) Using a pocket calculator and starting with the guess $x_0 = 2$, work out the first few iterates by hand and note how quickly this algorithm converges to $\sqrt{2} \approx 1.41421$. (This algorithm is the algorithm that most pocket calculators invoke when the square root key is pressed.) (c) Using whatever computational tool appeals to you, write a program that asks for the value of $a$, an initial guess for $\sqrt{a}$, and a tolerance and then implements Newton’s method to find $\sqrt{a}$, printing out each iterate along the way and stopping automatically when successive iterates differ by less than the specified tolerance.

14.13. For the first six Legendre polynomials $L_n(x)$, find all roots lying in the interval $-1 \leq x \leq 1$. Those polynomials are

$$
L_0(x) = 1 \\
L_1(x) = x \\
L_2(x) = \frac{1}{2}(3x^2 - 1) \\
L_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3) \\
L_5(x) = \frac{1}{4}(63x^5 - 70x^3 + 15x) \\
L_6(x) = \frac{1}{16}(231x^6 - 315x^4 + 105x^2 - 5)
$$

14.14. The natural frequencies for the transverse vibrations of a bar of uniform cross section that has length $L$ and is free at both ends are given by

$$
\omega_n = \frac{4K}{L^2} \sqrt{\frac{E}{\rho}} \alpha_n^2
$$

where $K$ is the radius of gyration of the cross section of the bar, $E$ is Young’s modulus for the material of the bar, $\rho$ is the density (mass/unit volume) of the material of the bar, and $\alpha_n$ is a solution to the equation

$$
tan \alpha = \pm tanh \alpha
$$


14.15. If the bar of the previous exercise is clamped at one end and free at the other, then the natural frequencies are given by the same expression except that $\alpha_n$ is instead a solution to the equation

$$
cot \alpha = \pm tanh \alpha
$$


Find the distinct roots of this polynomial and the multiplicity of each root.
14.16. The natural frequencies of the air in a spherical cavity are determined from the roots of the function $dj_n(x)/dx$, where $j_n(x)$ is the $n$-th order spherical Bessel function, the first three of which are

$$j_0(x) = \frac{\sin x}{x}; \quad j_1(x) = \frac{\sin x}{x^2} - \frac{\cos x}{x}; \quad j_2(x) = \left(\frac{3}{x^3} - \frac{1}{x}\right) \sin x - \frac{3}{x^2} \cos x$$

Obtain graphs of these three functions versus $x$, find the lowest several roots of $j_0(x)$, $j_1(x)$, and $j_2(x)$, and find also the lowest several roots of $dj_0(x)/dx$, $dj_1(x)/dx$, and $dj_2(x)/dx$.

14.17. Explore the way the energy levels of the well described in Section 14.1.5 change as the parameter $c$, which is determined by the depth and the width of the well, increases. At base, changing $c$ changes the slopes of the straight lines in Fig. 14.5. As $c$ increases and the well becomes deeper, the lines become more and more nearly horizontal and the number of energy levels increases. Seek ultimately to generate a graph that shows the energy of each allowed level on the vertical axis as a function of the parameter $c$ along the horizontal axis.

14.18. The intensity $I(x)$ in the diffraction pattern produced by a single slit is given by

$$\frac{I(x)}{I_0} = \frac{\sin^2 x}{x^2}$$

where $I_0$ is the intensity in the center and $x$ is related to the position of the observation point away from the central maximum. The zeroes in this pattern are easy to locate (they occur at $x = n\pi$, $n = 0, \pm 1, \pm 2, \ldots$). Careful location of the maxima, however, is more complicated. They don’t occur where $\sin^2 x = 1$ because of the influence of the denominator that steadily increases as $x$ increases. Locate the positions of the first half dozen maxima in this pattern, which—basically—is a request to find the roots of the derivative of the function (though note that not all roots correspond to maxima). Use at least three different methods and at least two different computational tools, and compare the results. Do your results confirm that the roots approach odd multiples of $\frac{1}{2}\pi$ as they become large? Optional: You might also find it interesting to approximate the function with a power series expansion for $\sin x$, keeping quite a few terms but converting the root finding problem into that of finding the roots of a polynomial. Then, use methods for finding roots of polynomials and see if you can come to understand how accuracy depends on how many terms you keep and which root you seek.

14.19. Using at least three different methods and at least two different computational tools, find the first half dozen roots of the zeroth-order Bessel function $J_0(x)$. Note that these roots are related to the radii of circular nodes in some of the vibrations of a circular membrane. The values of these roots tabulated in Abramowitz and Stegun\textsuperscript{32} are 2.404825577, 5.5200781103, 8.6537279129, 11.791544391, 14.9309177086, 18.0710639679. \textit{Hint:} Most computational tools have built-in capabilities for evaluating the Bessel functions. Consult the appropriate vendor manuals.

14.20. Suppose a particle moves in one dimension under the influence of the potential energy

$$V(x) = -\frac{V_0 a^2 (a^2 + x^2)}{8a^4 + x^4} \quad \Rightarrow \quad \frac{V(x)}{V_0} = -\frac{(1 + \pi^2)}{8 + \pi^2}$$

where $\pi = x/a$. Using at least three different methods and at least two different computational tools, find the coordinates $\bar{x}$ of all turning points when the total energy $E$ of the particle is $E = -0.2V_0$ and also when the total energy is $E = -0.1V_0$. Optional: Obtain graphs of the position of each turning point as a function of particle energy over the allowed range of energies for bound states.

14.21. Generate a graph showing the turning points of the potential energy given by Eq. (14.3) as a function of the energy of the particle.

14.22. Suppose a straight railroad track of length 1 mile (5280 ft) is held absolutely immovable at its two ends. On a hot summer day, the track expands in length by 1 ft. If the track bows upward from the earth in a circular arc, how high above the earth will the track rise at its midpoint? Hint: The geometry of this exercise is shown in the accompanying figure, where \( l \) is the original length of the track, \( d \) is the rise at the center, and \( a \) is the length, \( R \) the radius, and \( \theta \) half the subtended angle of the circular arc. Thus, \( R\theta = a/2 \), \( \sin \theta = l/(2R) \), and \( R - d = R \cos \theta \). The task is to find \( \theta \) and \( R \) from the first two of these equations and then use the third to find \( d \). The only quantities known \( a \) priori are \( l \) and \( a \).

14.23. A particle of mass \( m \) moves in a potential energy given as a function of position by \( V(x) = V_0 \cosh(x/a) \). Because this potential energy is an even function of \( x \), the upper and lower turning points have the same absolute value but opposite sign. Find the upper turning point as a function of energy and generate a graph showing that turning point as a function of energy. Hint: Measure position in units of \( a \) and energy in units of \( V_0 \).

14.24. A particular problem—see Problem 3-19 in the fourth edition of Fluid Flow by Rolf H. Sabersky, Allan J. Acosta, Edward G. Hauptmann, and E. M. Gates (Prentice-Hall, Upper Saddle River, NJ, 1999)—in fluid flow leads to the need to find the roots of the fourth-order polynomial \( 12x^4 - 12x^3 + 4x - 1 \). Use graphical methods to find bounds on the roots and at least three different computational approaches to find all real roots of this polynomial.

14.27. The image of a distant object produced on a viewing screen by a small aperture is actually a diffraction pattern. When the aperture is a circle of diameter \( d \) and light from the object strikes the screen containing the aperture at normal incidence, the intensity in the diffraction pattern at angle \( \theta \) from the normal is given by

\[
I(\theta) = \frac{I_0}{2\pi d \sin \lambda} \left[ \frac{2J_1(\pi d \sin \theta/\lambda)}{\pi d \sin \theta/\lambda} \right]^2
\]

where \( I_0 \) is the intensity at the center of the pattern, \( \lambda \) is the wavelength of the light illuminating the aperture, and \( J_1(x) \) is the first order Bessel function. Using the Bessel function routine that is assuredly built in to your computational tool (see the appropriate manuals), find the angle \( \theta \) at which the first zero in the diffraction pattern lies, expressing that result as a multiple of \( \lambda/d \). The resulting angle expresses the angular separation of two nearby objects such that the maximum in the diffraction pattern produced by one lies on top of the first minimum away from the maximum of the other. That separation is universally taken as a measure of the resolution of the optical system creating the images, and the condition requiring this positioning of the two maxima is called the Rayleigh criterion. Hint: The angle will be small, so you can safely use the approximation \( \sin \theta \approx \theta \).

14.28. The Lennard-Jones potential energy \( V_{LJ} \), which is given in terms of the coordinate \( r \) by the expression

\[
V_{LJ} = 4\epsilon \left[ \frac{\sigma^{12}}{r^{12}} - \frac{\sigma^6}{r^6} \right]
\]

or

\[
\frac{V_{LJ}}{\epsilon} = 4 \left[ \frac{1}{r^{12}} - \frac{1}{r^6} \right]
\]

where \( \epsilon \) and \( \sigma \) are constants and \( \bar{r} = r/\sigma \), plays a prominent role in some theories of chemical bonding. Obtain a graph of this potential energy and then obtain graphs of the lower turning point and the upper turning point as functions of the energy of the system over the range of energies for which the particle experiencing the potential energy is bound in the associated potential well.

14.17.4 Finding More than One Unknown

14.33. In a global positioning system, the raw data from which the position is determined consists of distances from various reference points together with knowledge of the location of those reference points. In two dimensions, for example, we might try to locate our position \((x, y)\) in a plane from knowledge that we are a distance \( r_1 \) from the point \((x_1, y_1)\) and a distance \( r_2 \) from the point \((x_2, y_2)\).
CHAPTER 14. FINDING ROOTS

Not all values we might assign are physically meaningful. For example, there is no point that is simultaneously 20 miles from point 1 and 30 miles from point 2 if points 1 and 2 are, in fact, 100 miles apart. Depending on the circumstances, there may be no points, one point, or—most often—two. Develop an algorithm for finding your location in two dimensions when you know your distance from each of two reference points whose coordinates are known, implement your algorithm in a program using whatever computational tool seems appropriate, solve two or three test problems that you invent, and—in particular—try to describe and defend the conditions under which two, one, or no solutions exist.

Optional: Extend your entire consideration into three dimensions, which will require knowledge of distance from each of (at least) three known locations. Hint: Quite a bit of information is available on the website www.trimble.com/gps. There is also an article in the January, 1994, issue of Physics Today (“Where I Stand” by Daniel Kleppner, page 9).

14.34. Given the three points \((x_i, y_i)\), \(i = 1, 2, 3\), (a) find symbolic expressions for the coefficients \(a\), \(b\) and \(c\) of the parabola \(y = ax^2 + bx + c\) that passes through these three points and then (b) find a symbolic expression for the value of \(x\) at which the extremum point of the parabola occurs. Finally, (c) determine numerically the angle at which the maximum range of a projectile occurs if the ranges at \(\theta = 39^\circ\), 40\(^\circ\), and 41\(^\circ\) are 0.7251744, 0.7259484, and 0.7258887, respectively.

14.35. Kirchhoff’s laws in DC circuit theory contend that the net current flowing into any node must be zero and that the net voltage drop around any closed path in the circuit must be zero. Remembering that the voltage drop \(\Delta V\) across a resistor \(r\) carrying current \(i\) is given by \(\Delta V = ir\) and using the symbols defined in Fig. 14.16, apply these laws to each of the circuits in the figure to generate a set of simultaneous, inhomogeneous linear equations for the unknown currents. Then, using symbolic methods solve each case for the unknown currents. Finally, determine the effective resistance defined by \(R_{\text{eff}} = V/I\) for each circuit. Assume that all batteries and resistors have known values and that quantities represented in the figures with the same symbol have the same value. Warning: For even simple circuits, Kirchhoff’s laws provide more equations than unknowns. Correctly written, these equations are guaranteed to be consistent. The subset to be solved, however, must be carefully chosen to make sure its members are linearly independent of one another.

14.36. The file \$HEAD/data/freefall.dat\) contains 31 lines, the \(i\)-th of which contains one numerical value—the value of the position \(x_i\) in cm of a particle at time \(t_i = (i - 1)/60\) s. You have reason to believe that the set of data \((t_i, x_i)\) with \(i = 1, 2, 3, \ldots, n\) is described by the parabolic relationship

\[
x = at^2 + bt + c
\]

The method of least squares identifies the optimum values of the coefficients \(a\), \(b\), and \(c\) as the particular values that minimize the residual

\[
R(a, b, c) = \sum_{i=1}^{n} \left( x_i - (at_i^2 + bt_i + c) \right)^2
\]

Using whatever language you choose, write a program that reads the positions from the file into a vector, generates a second vector containing the corresponding times and then enters a loop in which it asks for entry of trial values of \(a\), \(b\), and \(c\), calculates and displays \(R(a, b, c)\), and returns to ask for a new set of trial values. Using this program, conduct a manual search for the values of \(a\), \(b\), and \(c\) which make \(R(a, b, c)\) as small as possible. Hint: Think carefully about the initial guesses for \(a\), \(b\), and \(c\). A graph of \(x\) versus \(t\) may be useful.

14.37. The file \$HEAD/data/freefall.dat\) contains 31 lines, the \(i\)-th of which contains one numerical value—the value of the position \(x_i\) in cm of a particle at time \(t_i = (i - 1)/60\) s. You have reason to believe that the set of data \((t_i, x_i)\) with \(i = 1, 2, 3, \ldots, n\) is described by the parabolic relationship

\[
x = at^2 + bt + c
\]

The method of least squares identifies the optimum values of the coefficients \(a\), \(b\), and \(c\) as the particular values that minimize the residual

\[
R(a, b, c) = \sum_{i=1}^{n} \left( x_i - (at_i^2 + bt_i + c) \right)^2
\]
i.e., as the particular values satisfying the three equations
\[ \frac{\partial R}{\partial a} = 0 \quad ; \quad \frac{\partial R}{\partial b} = 0 \quad ; \quad \frac{\partial R}{\partial c} = 0 \]

These equations will turn out to be linear in \(a\), \(b\), and \(c\), with coefficients and inhomogeneities determined by sums of various products of the measured independent and dependent variables. Derive the three equations symbolically but then find the numerical values of the coefficients and inhomogeneities for the data in the file `data/freefall.dat` and use at least two different numerical approaches to find the solution for \(a\), \(b\), and \(c\). Finally, generate a graph in which each measured point is represented by a simple symbol and the least squares parabola is shown by a solid line so you can judge the adequacy of your fit.

14.38. A particle of mass \(m\) moves along the \(x\) axis under the action of a time-dependent force \(F(t)\). We observe that \(x(0) = 0\) and \(x(t_f) = a\). The detailed motion therefore is described by the solution to the boundary value problem
\[ m\frac{d^2x}{dt^2} = F(t) \quad , \quad x(0) = 0 \quad , \quad x(t_f) = a \]

To predict the detailed motion numerically, we might divide the interval \(0 \leq t \leq t_f\) into \(n\) segments of size \(\Delta t = t_f/n\), let \(x_i = i\Delta t\) with \(i = 0, 1, 2, \ldots, n\), evaluate the differential equation at \(t = t_i\), and introduce a finite difference approximation for the derivative to conclude that
\[ x_{i-1} - 2x_i + x_{i+1} = \frac{F(t_i)}{m}\Delta t^2 \]
which is valid for \(i = 1, 2, \ldots, n-1\). For the two end points, we remember the boundary values and require that
\[ x_0 = 0 \quad , \quad x_n = a \]
In total, we have \( n + 1 \) equations determining the \( n + 1 \) unknowns \( x_0, x_1, x_2, \ldots, x_n \). Cast these equations in the matrix form

\[
\begin{pmatrix}
? & ? & ? & \ldots & ? \\
? & ? & ? & \ldots & ? \\
? & ? & ? & \ldots & ? \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
? & ? & ? & \ldots & ?
\end{pmatrix}
\begin{pmatrix}
x_0 \\
x_1 \\
x_2 \\
\vdots \\
x_n
\end{pmatrix} =
\begin{pmatrix}
? \\
? \\
? \\
\vdots \\
?
\end{pmatrix}
\]

and note that the coefficient matrix is tridiagonal (and symmetric). Then, using at least two different computational tools, design and describe a general approach and implement that approach to determine the detailed motion for the cases

(a) \( t_f = 1 \) s, \( a = 10 \) m, \( F(t)/m = 8 \) N/kg (constant force), and \( n = 10 \);
(b) same as (a) but with \( n = 100 \);
(c) \( t_f = 1 \) s, \( a = 10 \) m, \( F(t)/m = e^{-t} \) N/kg (exponentially decaying force; \( t \) in seconds), and \( n = 10 \); and
(d) same as (c) but with \( n = 100 \).

Verify that the exact analytic solution in the two cases is

(a) and (b): \( x(t) = 4t^2 + 6t \); (c) and (d): \( x(t) = e^{-t} - 1 + (11 - e)t \)

and compare the numerical results with the exact solution in each case. (To save you a hunt, remember that the numerical value of \( e \) is 2.71828459045.)

14.39. The electrostatic potential on the boundary of the square region in a plane shown in Fig. 14.17 is maintained at the value zero along its bottom and right edges but increases linearly from zero along the left edge, then decreases linearly back to zero along the top edge. The potential in the interior region satisfies Laplace’s equation. One approach to finding an approximate solution involves imposing a regular grid on the region and then requiring that the value of the potential at each interior point be the average of the values at its four nearest neighbors, e.g.,

\[
V_{12} = \frac{1}{4} \left( V_{11} + V_7 + V_{13} + V_{17} \right) ; \quad V_{15} = \frac{1}{4} \left( V_{14} + V_{10} + 0.0 + V_{20} \right)
\]

(a) Write out the twenty-five equations implied by this requirement, (b) cast them in the matrix form

\[
\begin{pmatrix}
? & ? & ? & \ldots & ? \\
? & ? & ? & \ldots & ? \\
? & ? & ? & \ldots & ? \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
? & ? & ? & \ldots & ?
\end{pmatrix}
\begin{pmatrix}
V_1 \\
V_2 \\
V_3 \\
\vdots \\
V_{25}
\end{pmatrix} =
\begin{pmatrix}
? \\
? \\
? \\
\vdots \\
?
\end{pmatrix}
\]

figuring out the value represented by each question mark, and (c) solve the system for the unknowns. Then (d) structure the boundary values and solution in a \( 7 \times 7 \) matrix and create a surface plot of that solution.

14.40. In Ruchardt’s experiment, a steel ball bounces up and down in a vertical tube that ends in a gallon (or larger) jug. Ultimately, the ball falls into the jug, but it may bounce up and down many times before doing so. The file \$HEAD/data/ruchardt.dat contains 170 lines, the \( i \)-th of which contains one numerical value—the position \( x_i \) in cm of the steel ball at time \( t_i = 0.05(i - 1) \) s. A quick graph of the data (you should make it) suggests that the motion might be described by an exponentially decaying cosine curve on which is superimposed a linear sinking of the “equilibrium” position, i.e., by a function of the form

\[
x(t) = Ae^{-bt} \cos(\omega t + \phi) + Ct + D
\]
where the parameters $b$, $A$, $\omega$, $\phi$, $C$, and $D$ are to be determined. Develop a means to find optimal values of these six parameters by seeking values that minimize the quantity

$$ R = \sum_{i=1}^{N} \left( x_i - Ae^{-bt_i} \cos(\omega t_i + \phi) + Ct_i + D \right)^2 $$

**Warning:** This exercise is not for the faint-hearted. Some useful background will be found in Chapter 8 of *Data Reduction and Error Analysis for the Physical Sciences* (Second Edition) by Philip R. Bevington and D. Keith Robinson (McGraw-Hill, New York, 1992).
14.C Listing of OCTAVE Function lubisect.m

function root = lubisect(func, xlb, xub, tol, itmax)
# LUBISECT - Finds roots by method of bisection
# Function lubisect finds a root of the function func
# that lies between xlb and xub, using the method
# of bisection. The interval is repeatedly halved
# until |xub-xlb|<tol or itmax refinements have
# been made.

fl = feval( func, xlb );
fu = feval( func, xub );
fm = 1.0;

if fl*fu > 0.0
    root = xlb
    error('The given interval does not contain a root.')
endif

i = 0;
while (xub-xlb >= tol) && (fm ~= 0.0) && (i < itmax)
    xm = 0.5*(xlb + xub);
    fm = feval( func, xm );
    if fl*fm >= 0 xlb = xm; else xub = xm; endif
    i = i + 1;
endwhile

fprintf('
Number of iterations = %d
', i)
if i >= itmax fprintf('Convergence not achieved.
'); endif

root = xm;
endfunction
Chapter 15

Solving Partial Differential Equations

The properties—temperature, electrostatic potential, magnetic field, membrane displacements, quantum wave functions, fluid pressure, ...—of physical systems are usually encapsulated mathematically in functions of space and time. As the properties are constrained physically by the systems themselves, the functions representing the properties are constrained mathematically by various physical laws, which typically specify one or more relationships among the partial derivatives of the dependent variable(s) with respect to two or more independent variables and hence will take the form of one or more partial differential equations (PDEs). The important PDEs of mathematical physics are second-order equations and are frequently—though certainly not always—linear. In this chapter, we limit ourselves to linear equations.

To define a problem completely (i.e., so that it has a unique solution), the associated PDEs must be supplemented with appropriate boundary conditions, which will usually fall into one of three categories, specifically,

- **Dirichlet** boundary conditions or boundary conditions of the *first* kind, which specify values that the *solution* must assume on the boundary of the region in which a solution is sought,
- **Neumann** boundary conditions or boundary conditions of the *second* kind, which specify values that a *derivative* of the solution must assume on the boundary of that region, and
- **mixed** boundary conditions or boundary conditions of the *third* kind, which specify values that some linear combination of the solution and its derivatives must assume on the boundary of that region.

When time is among the independent variables, complete specification of a problem will also entail stipulation of appropriate initial conditions, which specify values that the function and—for equations that are second order in time—its time derivative must assume throughout the spatial domain at a specific time (usually taken to be time 0). Depending on which conditions are necessary, we face a boundary-value problem (BVP) or an initial-value problem (IVP) or, sometimes, a problem involving both boundary and initial conditions.

Although closed form, symbolic solutions exist for some BVPs and IVPs, most of the time such problems can only be solved approximately by numerical methods. Most commonly, approaches to the numerical solution of PDEs involve (1) selecting a (usually large) set of points or *nodes* (or, in some contexts, *vertices*) that cover the ranges of all (or all save one) of the independent variables, (2) approximating the PDE(s) in such a way as to convert it (them) into a (usually large) set of algebraic equations (AEs) or *ordinary* differential equations (ODEs), and (3) solving the resulting set
of AEs or ODEs for the (approximate) solution to the original problem at each node. For example, rather than seeking the solution \( u(x,y,t) \) to the diffusion equation in two dimensions, we might introduce a grid or mesh, i.e., a set of discrete points \((x_i,y_j,t_k)\) distributed somehow over the ranges of the variables, and then convert the PDE into a set of algebraic equations for the several values \( u_{i,j,k} = u(x_i,y_j,t_k) \). Alternatively, we could discretize only the spatial variables by introducing a grid defined by the points \((x_i,y_j)\) and convert the PDE into a set of ODEs for the several functions \( u_{i,j}(t) = u(x_i,y_j,t) \) of the continuous variable \( t \). In either case, the boundary and initial conditions constrain the values and/or derivatives of the solution at nodes on the boundary. Thus, the number of independent AEs or ODEs will turn out to be exactly equal to the number of unknown values or functions in the problem.

At least two distinctly different approaches to the conversion of a PDE into a set of AEs or ODEs are in common use. Finite difference methods (FDMs) are quick, (comparatively) easy to motivate, and fast to code in computer languages, but they are borderline impossible to apply unless (1) the nodes are uniformly spaced and (2) the boundaries of the region in which a solution is sought coincide with the coordinate lines in one of the standard coordinate systems (Cartesian, polar, cylindrical, spherical, ...). An algebraic expression, which references two or more adjacent nodes, is used to approximate each derivative in the differential equation at each of these nodes. Ultimately, the partial differential equation(s) is (are) replaced by a system of AEs, (or, in some cases, by a system of ODEs), which is then solved for the dependent variable at each node.

In finite element methods (FEMs), however, the nodes forming the grid may be non-uniformly spaced. Collections of these nodes form geometric shapes—lines in one dimension; triangles or quadrilaterals or ... in two dimensions; tetrahedrons or hexahedrons (sometimes called bricks) or ... in three dimensions—that divide the region of interest into subregions, i.e., into the (finite) elements that give the method its name. Interpolation or shape functions are defined to facilitate approximating the dependent variable at points within each element from presumed known values of the dependent variable at the nodes or vertices associated with that element. The differential equation is then replaced with an equivalent integral statement, which is in turn converted into a system of algebraic equations by substituting the shape functions into this integral form, integrating, assembling the results from all elements, and imposing the constraints dictated by the boundary and initial conditions. Thus, as with the finite difference approach, the original partial differential equation(s) is (are) replaced by a system of algebraic equations, which is then solved for the dependent variable at each node.

Finite element methods have significant advantages over finite difference methods. For example, since finite element methods allow a nonuniform mesh, portions of the region in which a solution is sought can be treated with a fine mesh while, at the same time, other portions of that region can be represented with a coarse mesh, so the computational effort can be focused where it is most needed. Furthermore, especially if the boundaries are irregular, the boundary conditions are more easily incorporated with finite element methods than with finite difference methods. While we gain considerably in generality by invoking a method of this second type, we pay a heavy price: finite element methods are difficult to describe and even more difficult to code.

In Section 15.1, we deduce and provide physical contexts for a number of important PDEs. Then, in subsequent sections, we introduce both finite difference and finite element methods, illustrating each with explicit solutions to representative problems. For the sake of the quickest exposition of the essential ideas, we start by addressing problems with one independent variable before moving on to problems having two or more independent variables and requiring coordinate systems other than Cartesian. Because the methods are elaborate and their coding in computer languages is involved and lengthy, this chapter will undoubtedly seem more mathematical and abstract than physical. However obscured it may appear to be by the detailed discussion of method and implementation, the desire to address physical situations does assuredly lie underneath the entire exposition.
15.1 Sample Problems

We begin by laying out several of the most common PDEs of mathematical physics, placing each in at least some of the physical contexts in which it appears. As we shall discover in Section 15.1.5, the most common equations fall into one of three categories, represented respectively by the (classical) wave equation, the diffusion equation, and the Laplace equation.

15.1.1 Motion of a String: The Wave Equation

Consider a string that extends along the (horizontal) $x$ axis when it is in its equilibrium position, and let a point of the string when the string is in that position be located at $x$. Suppose that the string is moving in a plane (not the most complicated possible motion) such that, at a general time $t$, the point nominally located at $x$ is displaced transversely (i.e., perpendicular to the equilibrium orientation of the string) by an amount $u(x,t)$ and longitudinally (i.e., parallel to that equilibrium orientation) by an amount $v(x,t)$. The geometry is shown in Fig. 15.1. The string is, of course, under tension $\tau(x,t)$, which in general will vary from point to point and from time to time but, for a perfectly flexible string (which we explicitly assume), will always be directed tangent to the string. A force diagram for an isolated element—the element located between $x$ and $x + \Delta x$ when the string is in equilibrium—of the string might look like Fig. 15.2. For the transverse motion of the string, Newton’s second law $F = ma$ then requires that

$$\rho(x) \Delta x \frac{\partial^2 u}{\partial t^2} = \tau(x + \Delta x, t) \sin \theta(x + \Delta x, t) - \tau(x, t) \sin \theta(x, t) - \rho(x) \Delta x g$$  \hspace{1cm} (15.1)

or, if we divide by $\Delta x$ and let $\Delta x$ approach zero, that

$$\rho(x) \frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x} \left( \tau(x, t) \sin \theta(x, t) \right) - \rho(x) g$$ \hspace{1cm} (15.2)

1\ We take this transverse direction to be vertical, and include gravity among the forces acting on the string.
Here, $\rho(x) \Delta x$ is the mass of the element, and $\rho(x) g \Delta x$—the gravitational force on the element when the string is near the surface of the earth—illustrates a possible external force on the string. Similarly, for the longitudinal motion of the string, Newton’s second law requires that

$$\rho(x) \Delta x \frac{\partial^2 v}{\partial t^2} = \tau(x + \Delta x, t) \cos \theta(x + \Delta x, t) - \tau(x, t) \cos \theta(x, t)$$

or that

$$\rho(x) \frac{\partial^2 v}{\partial t^2} = \frac{\partial}{\partial x} \left( \tau(x, t) \cos \theta(x, t) \right)$$

for the horizontal motion.

This set of equations still contains too many unknowns. We must eliminate $\theta(x, t)$, because we really want to find only $u(x, t)$ and $v(x, t)$. Figure 15.3 supports the conclusion that

$$\sin \theta = \frac{\Delta u}{\Delta s} = \frac{\Delta u}{\sqrt{\Delta u^2 + (\Delta x + \Delta v)^2}} = \frac{\partial u}{\partial x} \left[ \left( \frac{\partial u}{\partial x} \right)^2 + \left( \frac{1 + \frac{\partial v}{\partial x}}{2} \right)^2 \right]^{1/2}$$

and that

$$\cos \theta = \frac{\Delta x + \Delta v}{\Delta s} = \frac{1 + \frac{\partial v}{\partial x}}{\sqrt{\left( \frac{\partial u}{\partial x} \right)^2 + \left( 1 + \frac{\partial v}{\partial x} \right)^2}}$$

Thus, the equations describing the motion of the string are

$$\rho(x) \frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x} \left[ \tau(x, t) \frac{\partial u}{\partial x} \left[ \left( \frac{\partial u}{\partial x} \right)^2 + \left( 1 + \frac{\partial v}{\partial x} \right)^2 \right]^{1/2} \right] - \rho g$$

and

$$\rho(x) \frac{\partial^2 v}{\partial t^2} = \frac{\partial}{\partial x} \left[ \tau(x, t) \frac{1 + \frac{\partial v}{\partial x}}{\sqrt{\left( \frac{\partial u}{\partial x} \right)^2 + \left( 1 + \frac{\partial v}{\partial x} \right)^2}} \right]$$
The system is, of course, not complete as yet. We would also need to know not only appropriate initial conditions, i.e., initial values for the four functions

\[ u(x, 0), \quad v(x, 0), \quad \frac{\partial u}{\partial x}(x, 0), \quad \frac{\partial v}{\partial x}(x, 0) \quad (15.9) \]

but also appropriate boundary conditions on the string at its two ends and a connection between \( \tau \) on the one hand and \( u \) and \( v \) on the other. This system is decidedly non-linear and very difficult to solve.

Approximations are almost always necessary to turn the problem we would really like to solve into one that we can solve. If, for example, (1) the amplitude of the motion is small so that

\[ \Delta u \ll \Delta x \quad \Rightarrow \quad \frac{\partial u}{\partial x} \ll 1 \quad (15.10) \]

(2) \( \tau \) is sufficiently large that, in small amplitude motion, \( \tau \) remains essentially constant, and (3) the motion is transverse so that \( v = 0 \) everywhere and always, then the equation for \( v \) is automatically satisfied and the equation for \( u \) decouples from that for \( v \) and becomes the (inhomogeneous) wave equation,

\[ \rho \frac{\partial^2 u}{\partial t^2} = \tau \frac{\partial^2 u}{\partial x^2} - \rho g \quad (15.11) \]

which is correct under the given circumstances even if \( \rho \) varies with \( x \).

We have derived this result for motion of a string in one dimension. Two special cases are worth noting:

- If the situation is static so there is no time dependence (\( \partial u / \partial t = 0 \)), then the shape of a string hanging under its own weight is given by the solution to the equation

\[ \tau \frac{d^2 u}{dx^2} = \rho g \quad (15.12) \]

which reduces in this case to an ODE.

- If there is no outside force and \( \rho \) is constant, then Eq. (15.11) becomes

\[ \rho \frac{\partial^2 u}{\partial t^2} = \tau \frac{\partial^2 u}{\partial x^2} \quad \Rightarrow \quad \frac{\partial^2 u}{\partial t^2} = \frac{\tau}{\rho} \frac{\partial^2 u}{\partial x^2} \quad \Rightarrow \quad \frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} \quad (15.13) \]

Note that the constant \( c = \sqrt{\tau / \rho} \) has the units of velocity.\(^2\) It turns out—though we don’t know this as yet—that \( c \) is the speed of propagation of the wave conveyed by the solution to this equation.

In both cases, of course, appropriate boundary and initial conditions must be specified before the equation has a unique solution.

While its deduction is more complicated, Eq. (15.13) has a natural extension to two and three dimensions, in which case we would consider, for example, the displacement of a two-dimensional membrane or the pressure in a gas in a three-dimensional enclosure. The function we seek would then be \( u = u(x, y, z, t) = u(r, t) \) and the second derivative with respect to one spatial dimension becomes the Laplacian. The equation in that case is

\[ \frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} = \nabla^2 u \quad \text{or} \quad \left( \nabla^2 - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \right) = 0 \quad (15.14) \]

\(^2\)Since \( [\tau] = \text{N} = \text{kg} \cdot \text{m} / \text{s}^2 \) and \( [\rho] = \text{kg} / \text{m} \), \([\tau / \rho] = (\text{kg} \cdot \text{m} / \text{s}^2) / (\text{kg} / \text{m}) = \text{m}^2 / \text{s}^2 \!). (Here, the symbol \([\ldots]\) stands for the units of . . .)
Here
\[ \nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \]  
(15.15)
\[ = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2}{\partial \phi^2} + \frac{\partial^2}{\partial z^2} \]  
(15.16)
\[ = \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial}{\partial r} \right) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial}{\partial \theta} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2}{\partial \phi^2} \]  
(15.17)
where \((x, y, z), (r, \phi, z), \text{ and } (r, \theta, \phi)\) are the physicists' standard Cartesian, cylindrical, and spherical coordinates. We thus arrive at the (classical) wave equation, one of the three prototype equations in mathematical physics.

### 15.1.2 Heat Flow: The Diffusion Equation

To motivate a second of the important equations, we seek the temperature \(u(x, t)\) in the one-dimensional rod shown in Fig. 15.4. Let the rod be insulated along its sides and characterized by a thermal conductivity \(K\), which gives the rate of heat flow per unit area in the direction of increasing \(x\) by the expression
\[
\text{rate of heat flow in rod towards positive } x \text{ through cross section at right angles to rod} = -KA \frac{\partial u}{\partial x} \]  
(15.18)
where \(A\) is the area of a cross section of the rod. [In effect, Eq. (15.18) defines \(K\).] Here, the derivative \(\partial u/\partial x\) is the gradient of the temperature, and the minus sign appears because heat flows in the positive direction when the gradient is negative and in the negative direction when the gradient is positive, i.e., heat flows from regions of higher temperature to regions of lower temperature; the explicit minus sign then assures that \(K\) will be a positive quantity. With this definition of a material property, which may depend on position in the rod, we conclude that, as heat flows in the rod,
\[
\text{heat flow in time } \Delta t \text{ into shaded element across surface at } x = -KA \left. \frac{\partial u}{\partial x} \right|_x \Delta t \]  
(15.19)
and
\[
\text{heat flow in time } \Delta t \text{ into shaded element across surface at } x + \Delta x = +KA \left. \frac{\partial u}{\partial x} \right|_{x+\Delta x} \Delta t \]  
(15.20)
so that, in the end,
\[
\text{net heat conducted into shaded element in time } \Delta t = \left[ \left( KA \frac{\partial u}{\partial x} \right)_{x+\Delta x} - \left( KA \frac{\partial u}{\partial x} \right)_x \right] \Delta t \]  
(15.21)
This energy, of course, will affect the temperature of the element. Indeed, if we introduce two other properties of the material, its heat capacity per unit mass \( c \), and its density \( \rho \), we then conclude that

\[
\text{heat necessary to increase temperature of shaded element by amount } \Delta u = (A \Delta x \rho c) \Delta u \tag{15.22}
\]

where \( A \Delta x \) is the volume of the element, and (hence) \( \rho A \Delta x \) is its mass. Since any heat transported into the element must affect its temperature as per this equation, these two evaluations of the heat flux must be equal, and we conclude that

\[
\rho c \frac{\partial u}{\partial t} = \frac{\partial}{\partial x} \left( K \frac{\partial u}{\partial x} \right) \tag{15.23}
\]

or, if \( K \) happens to be constant, that

\[
\frac{\rho c}{K} \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \implies \frac{\partial^2 u}{\partial x^2} = \frac{1}{\alpha^2} \frac{\partial u}{\partial t} \tag{15.24}
\]

where \( \alpha^2 = K/\rho c \). The statement of the problem is, of course, not complete until appropriate initial and boundary conditions have been specified.

One special case is worth noting. If the situation is static, i.e., if thermal equilibrium has been reached, then \( \partial u / \partial t = 0 \) and the temperature distribution in the rod satisfies

\[
\frac{\alpha^2}{K} \frac{\partial u}{\partial x} = 0 \quad \text{or} \quad (K \text{ constant}) \quad \frac{\partial^2 u}{\partial x^2} = 0 \tag{15.25}
\]

with, of course, appropriate boundary values at the ends of the rod.

While its deduction is more complicated, Eq. (15.25) has a natural extension to two and three dimensions, in which case we would consider, for example, the evolution of the temperature distribution in a plate or in a three-dimensional object when the initial temperature throughout the object is given and either Dirichlet or Neumann or mixed boundary conditions are specified at all points on its boundary. We would then seek the function \( u(x, y, z, t) = u(r, t) \) giving the temperature at all points in space and time, and the fundamental equation becomes

\[
\nabla \cdot \left( K \nabla u \right) = \rho c \frac{\partial u}{\partial t} \quad \text{or} \quad (K \text{ constant}) \quad \nabla^2 u = \frac{1}{\alpha^2} \frac{\partial u}{\partial t} \tag{15.26}
\]

We thus arrive at the (classical) diffusion equation, the second of the three prototype equations in mathematical physics. The diffusion equation is similar to the wave equation, but differs from it in that the time derivative is only first order.

### 15.1.3 Steady State Heat Flow: The Laplace Equation

The third of the three important equations in mathematical physics is quickly deduced from Eq. (15.26). In two or three dimensions when a steady-state temperature distribution has been reached, \( u \) satisfies the Laplace equation

\[
\nabla^2 u = 0 \tag{15.27}
\]

which, as always, must be supplemented with appropriate boundary conditions before its solution is unique. (This time there will be no initial conditions, since there is no time variable in the picture.)

### 15.1.4 Other Situations

Variants on the three equations we have just deduced appear in many places. For example, Maxwell’s equations for the electrostatic field \( \mathbf{E} \) support the argument

\[
\nabla \times \mathbf{E} = 0 \quad \implies \quad \mathbf{E} = -\nabla V \\
\n\nabla \cdot \mathbf{D} = \nabla \cdot \epsilon \mathbf{E} = \rho \quad \implies \quad \nabla \cdot (\epsilon \nabla V) = -\rho \tag{15.28}
\]
leading to an equation that determines the electrostatic potential \( V \) from the charge density \( \rho \) and the dielectric permittivity \( \epsilon \). If \( \epsilon \) is constant, the equation becomes the Poisson equation,

\[
\nabla^2 V = -\frac{\rho}{\epsilon}
\]

(15.29)

If, in addition, \( \rho = 0 \), the equation becomes the Laplace equation

\[
\nabla^2 V = 0
\]

(15.30)

for the electrostatic potential.

For another example, suppose we seek a sinusoidal solution to the wave equation, i.e., suppose we seek a solution to Eq. (15.14) of the form

\[
u(x, y, z, t) = \psi(x, y, z) \cos \omega t
\]

Then, if \( u \) satisfies the wave equation and we set \( k^2 = \omega^2/c^2 \), \( \psi \) will satisfy the equation

\[
\nabla^2 \psi + k^2 \psi = 0
\]

an equation called the Helmholtz equation. In one dimension, the Helmholtz equation becomes

\[
\frac{d^2 \psi}{dx^2} + k^2 \psi = 0
\]

(15.31)

which describes standing waves in a string and many other physical situations.

We need not limit our enumeration to equations of importance in classical physics. For example, the time-dependent Schrödinger equation,

\[
-\frac{\hbar^2}{2m} \nabla^2 \psi + V \psi = i\hbar \frac{\partial \psi}{\partial t}
\]

(15.32)

for the quantum wave function \( \psi \) of a particle of mass \( m \) in a potential energy \( V \) is in some sense a diffusion equation (second order in the space derivatives, first order in the time derivative), though the term \( V \psi \) and the imaginary unit \( i = \sqrt{-1} \) create significant differences between the two equations.

Beyond Maxwell’s equations and the Schrödinger equation, second-order PDEs can be found in the Dirac equation (relativistic quantum mechanics), the equations of fluid dynamics (the Navier-Stokes’ equations; see Section 15.1.6), the equations of magnetohydrodynamics (MHD, which combine Maxwell’s equations and the Navier-Stokes’ equations), and in many other contexts in classical and contemporary physics.

### 15.1.5 Classification of Second-Order PDEs

In general terms, we have arrived in all cases at second-order, linear, often homogeneous PDEs, equations that we might collectively denote by the symbolism

\[
\mathcal{L} \{ u \} = 0
\]

(15.33)

where the operator \( \mathcal{L} \) symbolizes the (linear) differential operator that defines the equation. The most general example in two variables would have the form

\[
A(x, y) \frac{\partial^2 u}{\partial x^2} + 2B(x, y) \frac{\partial^2 u}{\partial x \partial y} + C(x, y) \frac{\partial^2 u}{\partial y^2} + D(x, y) \frac{\partial u}{\partial x} + E(x, y) \frac{\partial u}{\partial y} + F(x, y) u = 0
\]

(15.34)

where \( A, B, C, D, E, \) and \( F \) may depend on \( x \) and \( y \) but do not depend on \( u \) or its derivatives. Many, many important equations (though, significantly, not all) fall into this category. The solutions
of those equations that are in this category exhibit the very important property of superposition, namely
\[ L\{u_1\} = 0, \ L\{u_2\} = 0 \implies L\{au_1 + bu_2\} = aL\{u_1\} + bL\{u_2\} = 0 \] (15.35)
i.e., any linear combination of two solutions is itself a solution of the differential equation.

As it turns out, second-order, linear PDEs in two variables fall into three distinct categories, depending on the algebraic sign of the quantity \( AC - B^2 \). If, for example, \( AC - B^2 > 0 \), the equation is said to be elliptic; if \( AC - B^2 = 0 \), the equation is parabolic; and if \( AC - B^2 < 0 \), the equation is hyperbolic. When the coefficients actually depend on \( x \) and/or \( y \), an equation may fall into one category for some portions of the region of interest and another category for other portions of that region. Each of the prototype equations we have identified above, however, falls cleanly into one of the three categories. For the wave equation, we note that
\[ \frac{\partial^2 u}{\partial x^2} - \frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} = 0 \implies A = 1, B = 0, C = -\frac{1}{c^2} \implies AC - B^2 = -\frac{1}{c^2} < 0 \] (15.36)
and conclude that the wave equation is hyperbolic. For the diffusion equation,
\[ \frac{\partial^2 u}{\partial x^2} - \frac{1}{\alpha^2} \frac{\partial u}{\partial t} = 0 \implies A = 1, B = C = 0 \implies AC - B^2 = 0 \] (15.37)
and the diffusion equation is parabolic. Finally, for the Laplace equation,
\[ \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \implies A = C = 1, B = 0 \implies AC - B^2 = 1 \] (15.38)
and the Laplace equation is elliptic. In focusing on these three equations, we will therefore be exhibiting techniques that would be applicable to all possible second-order, linear, homogeneous PDEs in two variables.

Note, incidentally, that the variable transformation \( t \to iy \) will convert the wave equation (hyperbolic) into the Laplace equation (elliptic) and remember that this transformation also converts trigonometric functions into hyperbolic functions. Within the framework of complex variable theory, we might therefore expect that solutions to the wave equation and solutions to the Laplace equation could be related to one another (though we shall not pursue that connection in this book).

### 15.1.6 Equations of Fluid Dynamics: Navier-Stokes’ Equations

Let us next work out some of the fundamental equations of fluid dynamics. The state of a fluid at time \( t \) is described by a number of functions, the most important of which are the (vector) velocity field \( \mathbf{v}(\mathbf{r}, t) \), the (scalar) pressure \( p(\mathbf{r}, t) \), and the (scalar) density \( \rho(\mathbf{r}, t) \). In general, all of these quantities are functions of position \( \mathbf{r} \) within the fluid and of time \( t \). Further, \( p \) and \( \rho \) are related by the equation of state of the fluid involved.

Consider a volume—see Fig. 15.5—in the shape of a rectangular parallelepiped with its faces parallel to the coordinate planes. Let the lower left back corner be at \( (x, y, z) \) and the upper right front corner be at \( (x + \Delta x, y + \Delta y, z + \Delta z) \). We regard the volume as fixed in space, and the fluid as flowing through the volume in a way described by the velocity field already introduced. Further, the fluid in this volume experiences (internal) forces from its contact with the rest of the fluid at the surface of the volume and may also experience (external) forces from things like a nearby earth.\(^3\) The basic equations for fluid motion reflect the conservation of mass (the continuity equation), Newton’s second law, and conservation of linear and angular momentum. We deduce here only the first two of these fundamental relationships.

\(^3\)We assume the fluid is not electrically charged, and we ignore internal forces arising from the gravitational attraction of one portion of the fluid for another. Thus, the only internal forces on one element of the fluid will arise from its direct contact with contiguous elements.
15.1.6.1 Conservation of Mass: The Equation of Continuity

In the time interval from \( t \) to \( t + \Delta t \), the change in the mass of the fluid in the volume can be calculated in two different ways. First, we focus on the density, concluding that

\[
\text{mass added during interval } t \text{ to } t + \Delta t = \left[ \rho(t + \Delta t) - \rho(t) \right] \Delta x \Delta y \Delta z \approx \frac{\partial \rho}{\partial t} \Delta t \Delta x \Delta y \Delta z \tag{15.39}
\]

Alternatively, we can calculate this increment in mass from the velocity field. Focus for example on the two sides parallel to the \( yz \) plane. On the side at coordinate \( x \), all of the fluid in a volume \((v_x(x) \Delta t) \Delta y \Delta z\) passes through the surface into the volume in the time interval \( \Delta t \). Similarly, on the side at coordinate \( x + \Delta x \), all of the fluid in a volume \((v_x(x + \Delta x) \Delta t) \Delta y \Delta z\) passes out of the volume in that same time interval. On the first side, the density of the fluid is \( \rho(x) \) while on the second side the density is \( \rho(x + \Delta x) \). Thus, the net mass transported into the volume in time \( \Delta t \) over these two surfaces by the fluid flow is given by

\[
\text{net mass transported into volume through sides parallel to } yz \text{ plane in time } t \text{ to } t + \Delta t = \rho(x) v_x(x) \Delta t \Delta y \Delta z - \rho(x + \Delta x) v_x(x + \Delta x) \Delta t \Delta y \Delta z
\]

\[= - \left( \frac{\partial (\rho v_x)}{\partial x} \right) \Delta t \Delta x \Delta y \Delta z \tag{15.40}\]

All quantities in this last expression are evaluated at \( x, y, z, \) and \( t \). We justify ignoring differences in those variables from one point to another by noting that the expression is already first order in \( \Delta t, \Delta x, \Delta y, \) and \( \Delta z \). Therefore, any variation in \( \rho, v_x, \partial (\rho v_x)/\partial x, \) or \( \partial \rho/\partial t \) in the (small) interval from \( x \) to \( x + \Delta x \) would contribute at the strongest to second order in these small quantities—and we will consistently ignore contributions at that level.

Similar expressions apply for the other two pairs of faces (the pair parallel to the \( xy \) plane and the pair parallel to the \( xz \) plane). In total, the flow of the fluid through the volume transports a net mass given by

\[
\text{net mass transported into volume through all sides in time } t \text{ to } t + \Delta t = - \nabla \cdot (\rho v) \Delta t \Delta x \Delta y \Delta z
\]

\[= - \nabla \cdot (\rho v) \Delta t \Delta x \Delta y \Delta z \tag{15.41}\]
Figure 15.6: Stresses on the front (dotted) surface of an element in a fluid. (Be sure you visualize the intersection of the three vectors to lie on that front surface.) The signs are defined to give the stresses exerted on an element by the next element in the direction in which the coordinate increases.

The values obtained in Eqs. (15.39) and (15.41) for the mass added in the chosen volume in the time $\Delta t$ must, of course, be the same, and we arrive at the equation of continuity,

$$\nabla \cdot (\rho \mathbf{v}) + \frac{\partial \rho}{\partial t} = 0$$

(15.42)

which expresses the conservation of mass, i.e., the conviction that—however the fluid flows—any change in the mass in a fixed volume can come only by the flow of matter through the surface bounding that volume, i.e., that mass can be neither created nor destroyed in the volume. If the fluid is incompressible ($\rho$ constant), then the equation of continuity reduces to $\nabla \cdot \mathbf{v} = 0$.

### 15.1.6.2 Newton’s Second Law

The volume on which we are focusing our attention experiences forces from the surrounding fluid at its surface. Both normal forces (perpendicular to the surface and related to pressure) and shear forces (tangent to the surface and related to viscosity) may be experienced. All are expressed in terms of stress, normal stress for the first and shear stress for the second, where a stress in general is defined as a force per unit area. For stress, we use the notation $T_{rs}(x,y,z,t)$, where the first subscript conveys the direction of the normal to the surface involved and the second subscript conveys the direction of the stress, i.e.,

$$T_{rs} = \begin{pmatrix} \text{the component in the } s \text{ direction of} \\ \text{the force on a surface of unit area} \\ \text{oriented perpendicular to the } r \text{ direction} \end{pmatrix}$$

(15.43)

Even more specifically, $T_{xx}$ stands for the force in the (positive) $x$ direction (second subscript) on a surface of unit area oriented perpendicular to the $x$ axis (first subscript) and would be a normal stress. Similarly, the symbol $T_{xy}$ stands for the force in the (positive) $y$ direction (second subscript)

---

4The equation, of course, is quite similar to the parallel equation in electrodynamics expressing the conservation of charge.
on a surface of unit area oriented perpendicular to the \(x\) axis (first subscript) and would be a shear stress. The three stresses \(T_{xx}, T_{xy},\) and \(T_{xz}\) are shown in Fig. 15.6. As a vector, the total stress on a surface of unit area oriented perpendicular to the \(x\) axis would be given by

\[
\mathbf{T}_x = T_{xx} \hat{\mathbf{i}} + T_{xy} \hat{\mathbf{j}} + T_{xz} \hat{\mathbf{k}}
\]  

(15.44)

Note particularly that the stresses shown are the stresses on the indicated side of the illustrated element arising from its contact with the adjacent portion of the fluid in the direction in which the \(x\) coordinate increases. The stresses that the illustrated element exerts on its neighbor in the positive \(x\) direction will, via Newton's third law, be equal in magnitude but opposite in direction to those shown. In particular, then, the net force arising from internal stresses on the front and back surfaces of the illustrated element would be given by

\[
F_x = \left[ \left( T_{xx}(x + \Delta x, y, z) - T_{xx}(x, y, z) \right) \hat{\mathbf{i}} + \left( T_{xy}(x + \Delta x, y, z) - T_{xy}(x, y, z) \right) \hat{\mathbf{j}} \right] \Delta y \Delta z 
\]

\[= \left( \frac{\partial T_{xx}}{\partial x} \hat{\mathbf{i}} + \frac{\partial T_{xy}}{\partial x} \hat{\mathbf{j}} + \frac{\partial T_{xz}}{\partial x} \hat{\mathbf{k}} \right) \Delta x \Delta y \Delta z\]  

(15.45)

Here, we can evaluate the tensions at \(y, z\) because variation of \(y\) and \(z\) over the front and back surfaces will contribute only to second order in \(\Delta y\) and \(\Delta z\)—and we are ignoring terms beyond first order. Further, the derivatives can all be evaluated at argument \((x, y, z)\), since the variation between the front and back surfaces will contribute only to second order in \(\Delta x\). For clarity in the final result, those arguments have been omitted.\(^5\)

Only the subscripts and the variable with respect to which stresses are differentiated must be adjusted to yield expressions for the forces arising from internal stresses on the other two pairs of surfaces on the element in question. We find that

\[
F_y = \left( \frac{\partial T_{yx}}{\partial y} \hat{\mathbf{i}} + \frac{\partial T_{yy}}{\partial y} \hat{\mathbf{j}} + \frac{\partial T_{yz}}{\partial y} \hat{\mathbf{k}} \right) \Delta x \Delta y \Delta z
\]  

(15.46)

and

\[
F_z = \left( \frac{\partial T_{zx}}{\partial z} \hat{\mathbf{i}} + \frac{\partial T_{zy}}{\partial z} \hat{\mathbf{j}} + \frac{\partial T_{zz}}{\partial z} \hat{\mathbf{k}} \right) \Delta x \Delta y \Delta z
\]  

(15.47)

Finally, bringing together all of the \(x\) components of the forces arising from internal stress on all surfaces of the element on which our attention is focused and adding a possible external force\(^6\)

\[\rho f_x \Delta x \Delta y \Delta z,\]

we conclude that the \(x\) component of the net force on the element in question would be given by

\[
F_x = \left( \rho f_x + \frac{\partial T_{xx}}{\partial x} + \frac{\partial T_{yx}}{\partial y} + \frac{\partial T_{xz}}{\partial z} \right) \Delta x \Delta y \Delta z
\]  

(15.48)

According to Newton's second law, this force must, of course, also be given as the product of the mass of the element times the \(x\) component of its acceleration, i.e., by \((\rho \Delta x \Delta y \Delta z) a_x\). Thus, we have for the \(x\) component of the equation of motion the result

\[
\rho a_x = \rho f_x + \frac{\partial T_{xx}}{\partial x} + \frac{\partial T_{yx}}{\partial y} + \frac{\partial T_{xz}}{\partial z}
\]  

(15.49)

Similar equations could be deduced for the \(y\) and \(z\) components of the equation of motion for the fluid.

\(^5\)It is extremely easy to become very confused about these signs. Think about them several times.

\(^6\)For convenience, we define \(f_x\) to be a force per unit mass.
15.1. SAMPLE PROBLEMS

Evaluating the acceleration, however, introduces a subtle complication, since \( a_x \neq \frac{\partial v_x}{\partial t} \). This derivative is the rate of change of the velocity field at a fixed position in space. Unfortunately, at the end of the time interval, the element of the fluid on which the force acts is at a new position in the fluid. The quantity that figures in the definition of the acceleration we want is the change in the velocity of a particular element of the fluid that moves with the fluid; that quantity is given by

\[
\Delta \mathbf{v} = \mathbf{v}(x + v_x \Delta t, y + v_y \Delta t, z + v_z \Delta t, t + \Delta t) - \mathbf{v}(x, y, z, t)
\]

which implies that the acceleration we really want is given by

\[
a = \frac{\Delta \mathbf{v}}{\Delta t} = v_x \frac{\partial \mathbf{v}}{\partial x} + v_y \frac{\partial \mathbf{v}}{\partial y} + v_z \frac{\partial \mathbf{v}}{\partial z} + \frac{\partial \mathbf{v}}{\partial t} = \frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla) \mathbf{v}
\]

a derivative that is sometimes said to “follow the fluid” and is sometimes called the substantial derivative. Thus, the translation of the \( x \) component of Newton’s second law into the vocabulary used to describe the state of a fluid is

\[
\rho \left( \frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla \right) v_x = \rho f_x + \frac{\partial T_{xx}}{\partial x} + \frac{\partial T_{yx}}{\partial y} + \frac{\partial T_{zx}}{\partial z}
\]

Equation (15.52) and its \( y \) and \( z \) counterparts constitute a set of complicated, non-linear, coupled PDEs. Further, they are not by themselves complete, since they involve the stresses (which are not among the official quantities used to describe the state of the fluid). The equations (including the continuity equation) have too many unknowns (\( \rho, \mathbf{v}, \) and the stresses). We need some further relationships, particularly relationships linking stresses to fluid velocities (and involving new fluid parameters like viscosity). For example, in some simple cases, we can sometimes suppose that the shear stress in one direction (say \( x \)) on a surface perpendicular to another direction (say \( y \)) is related to the rate at which the component of the velocity in the first direction changes with displacement in the second direction, i.e., \( T_{yx} = \mu \frac{\partial v_x}{\partial y} \), where \( \mu \) is the viscosity of the fluid (which in effect is defined by this relationship).

15.1.6.3 A Special Case: Non-viscous Flow

If the fluid of interest has zero viscosity, then there can be no shear forces and all the off-diagonal elements of the stress tensor \( \mathbf{T} \) will be zero. Further, in most instances, the on-diagonal elements will all be equal to the negative of the pressure. That is, \( T_{xx} = T_{yy} = T_{zz} = -p \) and \( T_{yz} = T_{zy} = \ldots = 0 \). In that case, Eq. (15.52) reduces to

\[
\rho \left( \frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla \right) v_x = \rho f_x - \frac{\partial p}{\partial x}
\]

Combined in vector notation with the other two components, the basic equation for non-viscous flow then is that

\[
\rho \left( \frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla \right) \mathbf{v} = \rho \mathbf{f} - \nabla p
\]

Once the external force and the equation of state relating \( \rho \) and \( p \) have been determined, this equation provides the starting point for solving many problems in non-viscous fluid flow.

---

7To write a single equation combining the \( x, y, \) and \( z \) components, we would have to introduce the notation of tensors. Given our limited use of these relationships, we have little motivation to take that step.

8A fluid described by this relationship is said to be a Newtonian fluid. The relationship is an approximation, but—fortunately—there are several fluids to which it—and its generalizations to 2D and 3D flow—seem to be accurately applicable.
15.1.6.4 A Second Special Case: Sound Waves

Suppose the flow of interest is 1D, so \( v_y \) and \( v_z \) are both zero and \( v_x \) depends only on \( x \) and \( t \). Further, suppose that external forces are absent and that the flow is inviscid, i.e., that the viscosity is zero and hence that shear stresses are zero (\( T_{yx} = 0 \) and \( T_{zx} = 0 \)). Then Eq. (15.42) expressing conservation of mass and Eq. (15.52) expressing Newton’s second law become

\[
\frac{\partial}{\partial x} (\rho v_x) = -\frac{\partial \rho}{\partial x} \quad ; \quad \rho \left( \frac{\partial v_x}{\partial t} + v_x \frac{\partial v_x}{\partial x} \right) = -\frac{\partial p}{\partial x} \tag{15.55}
\]

(We have recognized that \( T_{xx} = -p \)). Further, suppose that the velocity of the fluid is small, so we can ignore the non-linear term and reduce these equations to

\[
\frac{\partial}{\partial x} (\rho v_x) = -\frac{\partial \rho}{\partial x} \quad ; \quad \rho \frac{\partial v_x}{\partial t} = -\frac{\partial p}{\partial x} \tag{15.56}
\]

Next, let us suppose that \( p = p(\rho) \), i.e., that the pressure in the fluid and its density are functionally related. Then,

\[
\frac{\partial p}{\partial x} = \frac{dp}{d\rho} \frac{\partial \rho}{\partial x} = K \frac{\partial \rho}{\partial x} \tag{15.57}
\]

where \( K = dp/d\rho \), which will in general depend on \( \rho \), is a property of the fluid characterizing the extent to which small changes in applied pressure are determined from small changes in density. The equations now become

\[
\frac{\partial}{\partial x} (\rho v_x) = -\frac{\partial \rho}{\partial x} \quad ; \quad \rho \frac{\partial v_x}{\partial t} = -K \frac{\partial \rho}{\partial x} \tag{15.58}
\]

Finally, we suppose that the pressure departs only slightly from its nominal equilibrium value \( \rho_0 \) and that the velocity in equilibrium is zero (the fluid is quiescent), i.e., we set

\[
\rho = \rho_0 + \delta \rho(x,t) \quad ; \quad v_x = \delta v_x(x,t) \tag{15.59}
\]

Then, ignoring all but the linear terms in small quantities, we find that the equations now are

\[
\rho_0 \frac{\partial \delta v_x}{\partial x} = -\frac{\partial \delta \rho}{\partial t} \quad ; \quad \rho_0 \frac{\partial \delta v_x}{\partial t} = -K_0 \frac{\partial \delta \rho}{\partial x} \tag{15.60}
\]

where \( K_0 = K(\rho_0) \) is the value of \( K \) associated with the equilibrium density \( \rho_0 \)—and is constant. Differentiating the first of these with respect to \( t \) and the second with respect to \( x \), we find that

\[
\rho_0 \frac{\partial^2 \delta v_x}{\partial t \partial x} = -\frac{\partial^2 \delta \rho}{\partial t^2} \quad ; \quad \rho_0 \frac{\partial^2 \delta v_x}{\partial x \partial t} = -K_0 \frac{\partial^2 \delta \rho}{\partial x^2} \tag{15.61}
\]

Since the mixed second-partial derivatives are equal, we then conclude that

\[
\frac{\partial^2 \delta \rho}{\partial t^2} = K_0 \frac{\partial^2 \delta \rho}{\partial x^2} \tag{15.62}
\]

That is, the density fluctuations \( \delta \rho \) in this system satisfy the wave equation. In effect, we have discovered that, among the motions possible in this medium, there is a longitudinal wave that propagates with speed \( v = \sqrt{K_0} \). [See Eq. (15.13) and footnote 2.]

As an aside, note that, for an ideal gas, \( pV = nRT \) or \( p = nRT/V = MRT/(Vm_n) \), where \( M \) is the mass of the sample of gas and \( m_n \) is the mass of the gas per mole. Further, when the gas undergoes an adiabatic process, as it does when it supports a sound wave, \( pV^\gamma \), where \( \gamma \) is the ratio of the specific heat at constant pressure to that at constant volume, is constant. Thus, since \( pV^\gamma = p_0V_0^\gamma \) and \( V = M/\rho \), we find that

\[
p \left( \frac{M}{\rho} \right)^\gamma = p_0V_0^\gamma \quad \Rightarrow \quad p = \frac{p_0V_0^\gamma}{M^\gamma \rho^\gamma} \quad \Rightarrow \quad K = \frac{dp}{d\rho} = \gamma \frac{p_0V_0^\gamma}{M^\gamma \rho^\gamma} \quad \rho^{-1} \tag{15.63}
\]
or, evaluating at $\rho = \rho_0$ and noting the relationship $V_0 = M/\rho_0$, that

$$K_0 = \frac{dp}{d\rho_0} = \gamma \frac{p_0V_0^2}{M^\gamma \rho_0^{\gamma - 1}} = \gamma \frac{p_0M^\gamma}{M^\gamma \rho_0^{\gamma - 1}} \rho_0^{-1} = \gamma \frac{p_0}{\rho_0} = \gamma \frac{RT_0}{m_n}$$  \hspace{1cm} (15.64)

and we have found $K_0$ for this simple system. Note that $K_0$ is temperature dependent. Further, with this value for $K_0$, we can predict that the speed of sound in an ideal gas at temperature $T_0$ is given by $v = \sqrt{K_0} = \sqrt{\gamma RT_0/m_n}$. In particular, we conclude that the speed of sound in a gas increases as the square root of the (absolute) temperature. With $R = 8.31 \text{ N m/(mol K)}$ and—for air—$m_n = 0.0288$ kg and $\gamma = 7/5 = 1.4$, we find at room temperature ($70 \degree\text{F} = 21 \degree\text{C} = 294 \text{ K}$) that

$$v = \sqrt{\frac{\gamma RT_0}{m_n}} = \sqrt{1.4 \frac{(8.31)(294)}{0.0288}} \text{ m/s} = 344 \text{ m/s}$$  \hspace{1cm} (15.65)

A further interesting connection appears if we remember from the equipartition theorem of kinetic theory that the rms speed $v_{rms}$ of molecules of mass $m$ in a gas at absolute temperature $T_0$ satisfies $\frac{1}{2}mv_{rms}^2 = \frac{3}{2}kT_0$ or, on multiplying by Avogadro’s number, $\frac{1}{2}m_nv_{rms}^2 = \frac{3}{2}RT_0$. We conclude that

$$v_{rms} = \sqrt{\frac{3RT_0}{m_n}} \implies v = \sqrt{\frac{\gamma RT_0}{m_n}} = \sqrt{\frac{\gamma}{3} \frac{3RT_0}{m_n}} = \sqrt{\frac{\gamma}{3} v_{rms}}$$  \hspace{1cm} (15.66)

We thus find that the speed of sound in a gas is a simple multiply—0.683 in the case of air—of the root mean square speed of the molecules of the gas and that both increase as the square root of the temperature!

### 15.1.7 A General 1D Equation

While a differential equation in one dimension is, of course, an ODE, not a PDE, description and illustration of the essential approaches underlying both finite difference methods and finite element methods are simplest in one dimension. We elect to begin there, seeking as an example to solve the general linear, second-order, self-adjoint,\(^9\) inhomogeneous equation

$$-\frac{d}{dx} \left( \alpha(x) \frac{d\varphi(x)}{dx} \right) + \beta(x) \varphi(x) = f(x) \quad \text{or} \quad -\alpha(x) \frac{d^2\varphi(x)}{dx^2} - \alpha'(x) \frac{d\varphi(x)}{dx} + \beta(x) \varphi(x) = f(x)$$  \hspace{1cm} (15.67)

in the interval $0 \leq x \leq L$. Here, $\varphi(x)$ is an unknown function, $\alpha(x)$ and $\beta(x)$ are parameters related to the physical properties of the problem (and may be functions of $x$), $\alpha'(x) = d\alpha(x)/dx$, and $f(x)$ is a forcing or source function (inhomogeneity). With appropriate choices of $\alpha$, $\beta$, and $f$, this equation can be reduced to the one-dimensional versions of several of the equations we have developed in the previous subsections.

To be complete, we shall suppose we are dealing with a boundary value problem, taking the desired boundary conditions to be given by

$$\varphi(0) = p$$  \hspace{1cm} (15.68)

and

$$\left. \left[ \alpha(x) \frac{d\varphi(x)}{dx} + \gamma \varphi(x) \right] \right|_{x = L} = q$$  \hspace{1cm} (15.69)

where $p$, $\gamma$, and $q$ are known constants. Together, these conditions will allow us to demonstrate how to treat each of the three kinds of boundary conditions. The condition at $x = 0$ is a Dirichlet boundary condition or boundary condition of the first kind; the condition at $x = L$ is a mixed boundary condition or boundary condition of the third kind. When $\gamma = 0$ in this second condition, the condition reduces to a Neumann condition or boundary condition of the second kind.

---

\(^9\)The word self-adjoint characterizes an equation in which—see the second form in Eq. (15.67)—the coefficient of the first derivative term is the derivative of the coefficient of the second derivative term. While this requirement appears to be restrictive, it turns out—see Exercise 15.3—that any linear, second-order equation can with an appropriately chosen integrating factor be cast in self-adjoint form, so the restriction is only apparent, not real.
15.1.8 A General 2D Equation

As a simple—and general—two-dimensional example, we will illustrate the application of finite difference and finite element analysis to solve an equation of the general form

\[-\frac{\partial}{\partial x} \left( \alpha_x(x, y) \frac{\partial \varphi}{\partial x} \right) - \frac{\partial}{\partial y} \left( \alpha_y(x, y) \frac{\partial \varphi}{\partial y} \right) + \beta(x, y) \varphi = f(x, y) \tag{15.70} \]

where \(\alpha_x, \alpha_y\), and \(\beta\) are known quantities—possibly constants; possibly functions of \(x\) and \(y\)—and \(f\)—also possibly constant; possibly a function of \(x\) and \(y\)—is a driving term (inhomogeneity). While this equation is presented in its most general form so that the results of our discussion can be applied to a variety of different physical problems, appropriate restrictions of \(\alpha_x, \alpha_y, \beta,\) and \(f\) will reduce it to the Laplace, Poisson, or Helmholtz equation. To complete the statement of the problem, we suppose that a solution is to be found subject to the Dirichlet conditions

\[\varphi = p \quad \text{(on } \Gamma_1)\] (15.71)

on the portion \(\Gamma_1\) of the boundary and the Neumann conditions

\[\left( \alpha_x \frac{\partial \varphi}{\partial x} i + \alpha_y \frac{\partial \varphi}{\partial y} j \right) \cdot \hat{n} = q \quad \text{(on } \Gamma_2)\] (15.72)

on the remainder \(\Gamma_2\) of the boundary. Here, \(p\) and \(q\) are quantities defined on the boundary, and \(\hat{n}\) is a unit vector perpendicular to the boundary and directed outward from the perspective of a viewer in the region in which a solution is sought.

15.2 Finite Difference Methods (FDMs) in One Dimension

In the finite difference approach to the one-dimensional problem laid out in Section 15.1.7, we begin by dividing the interval \(0 \leq x \leq L\) into \(N\) segments, each of length \(\Delta x = L/N\), so that the \(i\)-th node \((i = 0, 1, 2, \ldots, N)\) has \(x\) coordinate \(x_i = i \Delta x\). In particular \(x_0 = 0\) and \(x_N = L\). Then, we evaluate Eq. (15.67) at the point \(x_i\) to find that

\[-\alpha_i \frac{d^2 \varphi(x)}{dx^2} \bigg|_{x_i} - \alpha_i' \frac{d \varphi(x)}{dx} \bigg|_{x_i} + \beta_i \varphi_i = f_i \tag{15.73} \]

where \(\alpha_i = \alpha(x_i), \alpha_i' = \alpha'(x_i), \beta_i = \beta(x_i),\) and \(f_i = f(x_i)\) are all known and \(\varphi_i = \varphi(x_i)\) is to be found. Next, we approximate the derivatives by invoking finite differences. We illustrate only the most common of several possible ways to achieve that objective. In terms of the quantities \(\varphi_i,\) we might, for example, write the first derivative at \(x_i\) in any of the ways

\[\frac{d \varphi(x)}{dx} \bigg|_{x_i} \approx \frac{\varphi_{i+1} - \varphi_i}{\Delta x} \quad \text{or} \quad \frac{d \varphi(x)}{dx} \bigg|_{x_i} \approx \frac{\varphi_i - \varphi_{i-1}}{\Delta x} \quad \text{or} \quad \frac{d \varphi(x)}{dx} \bigg|_{x_i} \approx \frac{\varphi_{i+1} - 2 \varphi_i + \varphi_{i-1}}{2 \Delta x} \tag{15.74} \]

or probably in other ways as well (see Exercise 15.25). The first of these approximations involves a forward difference, the second involves a backward difference, and the third involves a central difference.\(^{10}\) Each is correct, though they are not all equally convenient or useful—nor is any single one always the most appropriate or convenient. For the present example, we choose the central difference to approximate the first derivative in Eq. (15.73). To find an approximation for the second derivative, however, we use both the forward and the backward approximations for the first derivative, writing that

\[\frac{d^2 \varphi(x)}{dx^2} \bigg|_{x_i} \approx \frac{d \varphi(x)}{dx} \bigg|_{x_i + \Delta x/2} - \frac{d \varphi(x)}{dx} \bigg|_{x_i - \Delta x/2} \quad \frac{\varphi_{i+1} - \varphi_i}{\Delta x} - \frac{\varphi_i - \varphi_{i-1}}{\Delta x} = \frac{\varphi_{i+1} - 2 \varphi_i + \varphi_{i-1}}{\Delta x^2} \tag{15.75} \]

\(^{10}\)Note that the central difference formula is the average of the forward and backward formulae.
Here, we have recognized that the forward (backward) difference approximation to the derivative at \( x_i \) is a central difference approximation to the derivative at \( x_i + \frac{1}{2} \Delta x \) \( (x_i - \frac{1}{2} \Delta x) \), and we have taken the second derivative at \( x_i \) to be approximated by the difference of these two approximations to the first derivative divided by the separation of the points at which the two first derivatives are evaluated.\(^{11}\) Finally, we substitute the approximation of Eq. (15.75) and the central difference approximation of Eq. (15.74) into Eq. (15.73) to find that

\[
-\alpha_i \left( \frac{\varphi_{i+1} - 2 \varphi_i + \varphi_{i-1}}{\Delta x^2} \right) - \alpha_i' \left( \frac{\varphi_{i+1} - \varphi_{i-1}}{2 \Delta x} \right) + \beta_i \varphi_i = f_i
\]

(15.76)

or, on multiplying by \( \Delta x^2 \) and collecting terms with the same index on \( \varphi \), that

\[
\left(-\alpha_i + \frac{\alpha_i' \Delta x}{2}\right) \varphi_{i-1} + \left(2 \alpha_i + \beta_i \Delta x^2\right) \varphi_i + \left(-\alpha_i - \frac{\alpha_i' \Delta x}{2}\right) \varphi_{i+1} = f_i \Delta x^2
\]

(15.77)

Since \( i \) can assume any value in the interval \( 0 \leq i \leq N \), we thus have \( N + 1 \) equations, just the right number to determine the \( N + 1 \) unknowns \( \varphi_i \).

The conclusion of the last sentence of the previous paragraph, however, is premature. Unfortunately, when evaluated at \( i = 0 \) or \( i = N \), Eq. (15.77) makes reference to \( \varphi_{-1} \) or \( \varphi_{N+1} \), each of which is outside the domain of the problem! Thus, we really have \( N + 3 \) unknowns. The additional information we need lies in the boundary conditions, though the precise way in which these conditions resolve this problem depends on the type of boundary condition. For the \textit{Dirichlet} condition of Eq. (15.68), we abandon the first equation \((i = 0)\) altogether, replacing it with the equation prescribed by the boundary conditions, namely

\[
\varphi_0 = p
\]

(15.78)

For the \textit{mixed} boundary condition of Eq. (15.69), we use a central difference approximation to the derivative in the boundary condition, finding that

\[
\alpha_N \left( \frac{\varphi_{N+1} - \varphi_{N-1}}{2 \Delta x} \right) + \gamma \varphi_N = q \quad \implies \quad \varphi_{N+1} = \varphi_{N-1} + \frac{2 \Delta x}{\alpha_N} (q - \gamma \varphi_N)
\]

(15.79)

where \( \alpha_N = \alpha(x_N) = \alpha(L) \). Then, we write out the last equation \((i = N)\) and substitute from Eq. (15.79) to eliminate \( \varphi_{N+1} \), finding ultimately that

\[
-2 \alpha_N \varphi_{N-1} + \left[2 \left( \alpha_N + \gamma \Delta x\right) + \left( \beta_N + \frac{\gamma \alpha_N'}{\alpha_N} \right) \Delta x^2 \right] \varphi_N = \left( f_N + \frac{\alpha_N' q}{\alpha_N} \right) \Delta x^2 + 2q \Delta x
\]

(15.80)

With these resolutions of the first and last equations, we arrive at the now fully defined and complete set

\[
\varphi_0 = p
\]

\[
\left(-\alpha_i + \frac{\alpha_i' \Delta x}{2}\right) \varphi_{i-1} + \left(2 \alpha_i + \beta_i \Delta x^2\right) \varphi_i + \left(-\alpha_i - \frac{\alpha_i' \Delta x}{2}\right) \varphi_{i+1} = f_i \Delta x^2, \quad 1 \leq i \leq N - 1
\]

\[
-2 \alpha_N \varphi_{N-1} + \left[2 \left( \alpha_N + \gamma \Delta x\right) + \left( \beta_N + \frac{\gamma \alpha_N'}{\alpha_N} \right) \Delta x^2 \right] \varphi_N = \left( f_N + \frac{\alpha_N' q}{\alpha_N} \right) \Delta x^2 + 2q \Delta x
\]

(15.81)

of \( N + 1 \) equations for the \( N + 1 \) unknowns \( \varphi_0, \varphi_1, \ldots, \varphi_N \). Remember that, for definiteness, we have chosen to use \textit{Dirichlet} boundary conditions at \( x = 0 \) and \textit{mixed} boundary conditions at \( x = L \). In other situations, the conditions at the two ends might both be of the same type, or they could each

\(^{11}\)You may have to read this sentence several times. The appearance of the binomial coefficients \((1, 2, 1)\), as in \((a + b)^2 = a^2 + 2ab + b^2\), is worth noting.
be of a type different from those here illustrated. In any case, we have illustrated how to address all three possible types in the specific choices made.

In effect, we have deduced a set of linear algebraic equations to be solved for the unknowns $\varphi_0, \varphi_1, \ldots, \varphi_N$. For visualization, we note that the set can be seen in matrix form. Further, since each equation involves no more than three consecutive indices, the matrix of the coefficients is tridiagonal though not necessarily symmetric. For example, if $N = 10$, the matrix version of these equations would have the general form

$$
\begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
? & ? & ? & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & ? & ? & ? & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & ? & ? & ? & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & ? & ? & ? & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & ? & ? & ? & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & ? & ? & ? & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & ? & ? & ? & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & ? & ? & ? \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & ? & ? \\
\end{pmatrix}
\begin{pmatrix}
\varphi_0 \\
\varphi_1 \\
\varphi_2 \\
\varphi_3 \\
\varphi_4 \\
\varphi_5 \\
\varphi_6 \\
\varphi_7 \\
\varphi_8 \\
\varphi_9 \\
\varphi_{10}
\end{pmatrix}
= 
\begin{pmatrix}
p \\
? \\
? \\
? \\
? \\
? \\
? \\
? \\
? \\
? \\
? \\
\end{pmatrix}
$$

(15.82)

In this display, only those elements marked with '?' can differ from zero (though some of those may be zero as well), and every non-zero element except those in the column of $\varphi_i$'s is known at the outset.

Two cases are important. In the first case, the equations are inhomogeneous and the determinant of the coefficient matrix is not zero; and the equations have a unique solution that will approximate the solution to the original boundary value problem. In the second case, the equations are homogeneous, and we hope that the the coefficient matrix contains a parameter that can be adjusted to make the determinant of that matrix zero; and our problem involves seeking the eigenvalues and eigenvectors of the coefficient matrix; we expect to find several solutions, one corresponding to each eigenvalue of the coefficient matrix.$^{12}$

Finite difference approaches to PDEs are, of course, subject to error. In general, these errors fall into three distinct and independent categories. In the first instance, we have replaced a continuous variable by a discrete set of values of that variable, i.e., we have discretized the independent variable. In so doing, we have rendered the solution vulnerable to discretization error, an error that will be reduced as the separation of the discrete values is made smaller. From that perspective alone, we would want to make that separation as small as possible, the smaller the better. Unfortunately, as we make the separation smaller, we also increase the computational labor of generating the solution and hence the time required to complete the solution. What’s worse, increasing the computational labor also increases the likelihood that errors will be generated by computer roundoff, which arises because computers retain floating point numbers only to finite precision. We thus must seek a compromise: We want to use a grid with small intervals so as to reduce discretization error but we can’t make the grid too small because roundoff errors may then become significant. Fortunately, in most cases, we will be able to find a grid that is simultaneously fine enough to render discretization error of little consequence and large enough to keep roundoff errors at bay. The standard assessment of these issues involves solving the problem twice, once with a coarser grid and then again with a finer grid. Until roundoff becomes a problem, we can (almost) always safely assume that the solution on the finer grid is more accurate than that on the coarser grid, and a comparison of the two solutions will provide some indication of the accuracy of the solution.

Sometimes, we will use a direct algorithm to solve the discretized equations that convey the solution, and the third type of error will not be present. When an iterative algorithm is used, $^{12}$Two other cases exist but are of no significance. The equations might be inhomogeneous and the determinant of the coefficient matrix zero, in which case no solution exists, or the equations might be homogeneous and the determinant of the coefficient matrix non-zero, in which case only the trivial solution (all $\varphi$'s zero) exists.
however, we need be concerned not only about discretization and roundoff errors; we must also assess convergence error. In those cases, we are using a method that may or may not converge to the solution of the approximate equations. We must therefore worry not only about the extent to which the solution to the equations we are solving actually represents the solution to the original problem (discretization and roundoff errors) but also about the extent to which the solution we ultimately deduce is actually a solution to the approximating equations (convergence error). Again, we assess the accuracy of the solution to the approximating equations by comparing successive iterates that we hope are converging on that solution. That hope will be more or less justified depending on the rapidity of the apparent convergence.

We can carry our analysis no further without selecting a particular programming language in which to implement explicit coding. That task is undertaken in the next section(s).

15.5 Using OCTAVE to Solve 1D PDEs via an FDM

15.5.1 A General Coding

The final step is to solve Eq. (15.82) for the unknowns \( \varphi_0, \varphi_1, \ldots, \varphi_N \). The numerical operations associated with constructing and then solving Eq. (15.82) are clearly a job for a computer, especially when the problem to be solved involves more than a very few simultaneous equations. A program to construct Eq. (15.82) and then solve the resulting system for \( \varphi \) at each node would begin by asking the user to input the values of the various parameters involved in the problem. In all cases, the parameters \( p, \gamma, \) and \( q \), which relate to the boundary conditions, are constants. In general, the quantities \( \alpha, \alpha', \) and \( \beta \), which relate to the physical system involved, and \( f \), which represents a source or excitation function, will depend on \( x \). For simplicity in illustration, however, we will suppose these four quantities to be constants as well.\(^{13}\) Indeed, if \( \alpha \) is constant, then \( \alpha' \) is zero, and we will simply leave out terms multiplied by \( \alpha' \) when we construct the equations.

The first segment of the program we wish to write will, then, request input of all the constants needed in the remainder of the program. Appropriate OCTAVE statements are

\[
\begin{align*}
N & \quad = \text{input}(\ 'Enter number of segments (N):' \ ); \\
alpha & \quad = \text{input}(\ 'Enter \alpha:' \ ); \\
beta & \quad = \text{input}(\ 'Enter \beta:' \ ); \\
f & \quad = \text{input}(\ 'Enter f:' \ ); \\
L & \quad = \text{input}(\ 'Enter L:' \ ); \\
p & \quad = \text{input}(\ 'Enter p:' \ ); \\
gamma & \quad = \text{input}(\ 'Enter \gamma:' \ ); \\
q & \quad = \text{input}(\ 'Enter q:' \ );
\end{align*}
\]

Next, with the statements

\[
\begin{align*}
dx & = L/N; \quad dx2 = dx^2; \\
x & = dx*[\ 0 : N ];
\end{align*}
\]

we calculate the size of each segment \( dx \), the square of that size \( dx2 \), and the values of \( x \) at which (equally spaced) nodes will be placed in the interval \( 0 \leq x \leq L \). Then, to prepare for using OCTAVE’s operator \( \backslash \) to solve the equations, we must create \( cf \), an \((N+1) \times (N+1)\) matrix containing the coefficients of the system to be solved, and \textit{inhomo}, a vector containing the inhomogenieties. We invoke the statements

\(^{13}\)If \( \alpha, \alpha', \beta, \) and \( f \) are not constant, each of these quantities would have to be represented by an \( N + 1 \) element vector and values would have to be given for all of those elements.
inhomo = zeros(N+1,1) + 1.0;
inhomo = f*dx2*inhomo;
inhomo(1) = p;
inhomo(N+1) = inhomo(N+1) + 2.0*q*dx;

cf = zeros(N+1,N+1);
for i=1:N+1 cf(i,i) = 2.0*alpha + beta*dx2; end

for i=1:N cf(i,i+1) = -alpha; end
for i=1:N cf(i+1,i)= -alpha; end

for i=1:N+1 cf(N+1,i)= -2.0*alpha;

to create the necessary matrix and vector, which reflect Eqs. (15.81) and (15.82). Finally, we exploit the single statement

phi = cf\inhomo;

to solve the system of equations defined by the several vectors.

A more fully commented command file containing these statements is named fdm1d.m, is listed in Appendix 15.A.3, and can be copied from the directory $HEAD/octave.14

15.5.2 An Example: Simple Harmonic Motion

We now illustrate the application of fdm1d.m to a simplification of Eq. (15.67), specifically the equation

\[ m \frac{d^2 \varphi}{dt^2} + k \varphi = 0 \quad ; \quad t \in (0,T) \]  

(15.83)

for the simple harmonic motion of a mass \( m \) attached to a spring of stiffness \( k \). To match this situation to that discussed in Section 15.1.7, we must interpret \( \alpha \) as a constant equal to \(-m\), \( \beta \) as a constant equal to \( k \), \( L \) as the time \( T \) at the end of the interval of interest, and the independent variable \( x \) as the independent variable \( t \). Further, we must set \( f = 0 \). The dependent variable \( \varphi \) gives the displacement of the oscillator from its equilibrium position. Basically, we are interested in the position over a range of times \( 0 \leq t \leq T \). In a consistent set of units, we will take

\[ m = -\alpha = 4.0 \text{ kg} \quad ; \quad k = \beta = 3.0 \text{ N/m} \]

To complete the definition of the problem, we need to specify the boundary conditions.\(^\text{15}\) Suppose we seek a solution for which \( \varphi(0) = 0 \text{ m} \) and \( d\varphi(t)/dt|_{t=T} = 1.0 \text{ m/s} \), i.e., we specify the position at \( t = 0 \) and the velocity at \( t = T \). To effect these conditions, we need to assign the values

\[ p = 0.0 \text{ m} \quad ; \quad \gamma = 0.0 \text{ kg/s} \quad ; \quad q = -4.0 \text{ kg\cdot m/s} \]

the first of which reduces Eq. (15.68) to \( \varphi(0) = 0 \text{ m} \) and the rest of which (with \( \alpha = -4.0 \text{ kg} \)) reduce Eq. (15.69) to \( d\varphi(t)/dt|_{t=T} = 1.0 \text{ m/s} \). In this example, we will seek a solution over the interval \( 0 \leq t \leq 10 \text{ s} \), so \( L \to T = 10.0 \text{ s} \). In physical terms, we start the oscillator at its equilibrium position with an unspecified initial velocity such that, at \( t = 10.0 \text{ s} \), its velocity will be \( 1.0 \text{ m/s} \).

\(^{14}\)See the Local Guide for the meaning of $HEAD at your site.

\(^{15}\)Note that we are using a method that requires one item of information at each end of the region of interest—a boundary value problem. More often in problems in motion, one has an initial value problem in which one specifies two items at one end of the region of interest, say an initial position and an initial velocity.
With these choices, we are now ready to execute the command file and solve the problem. The statement

\texttt{fdm1d}

will execute the statements in the file, one at a time. The first several statements request input, to which we respond with the values

\begin{verbatim}
Enter number of segments (N): 20
Enter alpha:       -4.0
Enter beta:        3.0
Enter f:           0.0
Enter L:           10.0
Enter p:           0.0
Enter gamma:       0.0
Enter q:           -4.0
\end{verbatim}

Once we have entered these parameters, the remaining statements will construct the matrix and generate the solution, storing it in an array named \texttt{phi}. We will have no further interaction with the program. Presently, the OCTAVE prompt will return. At that point, all variables to which values have been assigned in the execution of the command file will be accessible at OCTAVE’s main prompt. In particular, we could plot the solution by invoking the statements

\begin{verbatim}
plot( x, phi, 'Color', 'black', 'LineWidth', 4 )
title('N=20', 'FontSize', 20 )
grid on
\end{verbatim}

The resulting graph is shown in the upper left panel of Fig. 15.7.

To assess the accuracy of this solution, we repeat the process with an increasing number of divisions of the interval in which the solution is sought. To save the solution just generated and generate others for \( N = 50 \) and \( N = 100 \), we execute the statements

\begin{verbatim}
x020=x; phi020=phi;
fdm1d
Enter number of segments (N): 50
. (rest same as for N=20)
x050=x; phi050=phi;
fdm1d
Enter number of segments (N): 100
. (rest same as for N=20)
x100=x; phi100=phi;
\end{verbatim}

Then, to plot all solutions in a single display, we execute the statements

\begin{verbatim}
subplot(2,2,1);
plot( x020, phi020, 'Color', 'black', 'LineWidth', 2 )
title('N=20', 'FontSize', 16 )
grid on
subplot(2,2,2);
plot( x050, phi050, 'Color', 'black', 'LineWidth', 2 )
title('N=50', 'FontSize', 16 )
grid on
subplot(2,2,3);
\end{verbatim}
Figure 15.7: Harmonic motion via finite difference analysis. This graph was produced with OCTAVE.

\begin{verbatim}
plot( x100, phi100, 'Color', 'black', 'LineWidth', 2 )
title('N=100', 'FontSize', 16 )
grid on
subplot(2,2,4);
plot( x020, phi020, 'Color', 'black', 'LineWidth', 2 )
title('N=20, 50, 100', 'FontSize', 16 )
grid on
hold on
plot( x050, phi050, 'Color', 'black', 'LineWidth', 2 )
plot( x100, phi100, 'Color', 'black', 'LineWidth', 2 )
\end{verbatim}

All of these graphs are shown in Fig. ??.

Further, as revealed most directly in the lower right panel, even the solution with only 20 segments lies quite close to the solutions for 50 and 100 segments. We conclude that the solution we have obtained is accurate, at least to the resolution of the graph paper.

As a further test on the accuracy of the solution, which we expect to improve as the number of segments increases, we compare the solution for \( N = 50 \) with that for \( N = 100 \) with the statements

\begin{verbatim}
for i=1:50 dif(i) = phi100(2*i-1) - phi050(i); end
[ max(dif), min(dif) ]
ans =  0.0114  -0.0138
[ max(phi100), min(phi100) ]
ans =  1.5964  -1.5972
\end{verbatim}

which suggest that the solution for \( N = 50 \) is accurate to an absolute error of about \( \pm 0.012 \) and that the solution falls in the range \(-1.6 \leq \phi \leq 1.6\). If we generate a solution for \( N = 200 \), its difference from the solution for \( N = 100 \) ranges from \(-0.0036 \) to \( 0.0029 \). so the solution for \( N = 100 \) evidently has an absolute error of about \( \pm 0.004 \).
The above tests of accuracy have, of course, assessed only discretization error. To test for roundoff error, we would have to increase the number of nodes even further, looking for the point at which the solution begins to depart from that to which it appears at the moment to be converging. Since we have in this example used a direct method for finding the solution to the system of linear equations, we need not be concerned about convergence error.

Finally, using the solution for \( N = 100 \), we might determine the initial velocity and check the final velocity with the statements

\[
\text{InitialVelocity} = \frac{\text{phi}(2) - \text{phi}(1)}{x(2) - x(1)}
\]
\[
\text{InitialVelocity} = -1.3820
\]
\[
\text{FinalVelocity} = \frac{\text{phi}(101) - \text{phi}(100)}{x(101) - x(100)}
\]
\[
\text{FinalVelocity} = 0.9587
\]

(The results are, of course, in m/s.) Despite the graphical agreement of the solutions for \( N = 50 \) and \( N = 100 \), the final velocity for \( N = 100 \) doesn’t quite match the prescribed boundary value of 1.0 m/s. For comparison, had we used \( N = 200 \), we would have found the initial and final velocities to be \(-1.3847 \text{ m/s and } 0.9793 \text{ m/s, respectively; had we used } N = 500, \) we would have found those velocities to be \(-1.3855 \text{ and } 0.9917 \text{ m/s. Apparently, with increasing } N, \) the calculated final velocity converges on the prescribed boundary value.

### 15.9 Finite Element Methods (FEMs) in One Dimension

Finite element methods provide an alternative to finite difference methods for approximating the solution of a boundary value problem—ordinary or partial differential equation plus boundary conditions—in one or more dimensions. The total process involves

1. **Discretizing the domain (preprocessing).** The region of interest is subdivided into a number of small elements by appropriately chosen nodes, at each of which an approximation to the dependent variable \( \varphi \) will be sought. The discretization of the domain through the identification of suitable nodes and elements is known as preprocessing.

2. **Selecting the interpolation or shape functions.** The interpolation or shape functions for approximating the dependent variable within an element are selected. Because of its simplicity, linear interpolation is commonly used. Higher-order polynomials are more accurate but they also result in a more complicated formulation.

3. **Formulating the equations for a single element.** Equations for each element—the elemental equations—are formulated using the Ritz variational method, the Galerkin method, or some alternative and less common method.

4. **Assembling the system of equations.** The full system of equations is then obtained by (1) assembling the elemental equations into a set applying to the entire region and (2) imposing continuity conditions at the interfaces between elements.

5. **Incorporating the boundary conditions.** Constraints imposed at the boundaries of the region of interest are brought to bear to resolve unknowns left undetermined by the previous steps.

6. **Solving the system of equations.** The (probably large) system of simultaneous, linear, algebraic equations deduced at the previous step is solved for the dependent variable at each node. Such numerical methods as Gauss-Jordan elimination, Gaussian elimination with backsubstitution, and LU (lower-upper) decomposition can all be invoked.
Figure 15.8: Discretization of the solution domain in one dimension: (a) element and global node numbers; (b) element with local node numbers. As illustrated, the spacing of the nodes—and hence the lengths of the elements—need not be uniform.

7. Displaying the solution (postprocessing). The resulting solution is displayed in various ways. Separate from the finite element method, this last step is known as postprocessing. Postprocessing can be a totally separate process from the other steps, and may even be carried out with different software packages than were used to find the solution itself.

In this and the next section(s), we illustrate the method of finite element analysis by applying these steps to the one-dimensional boundary value problem defined in Section 15.1.7.

15.9.1 Discretizing the Domain: Preprocessing

The first step in the finite element method is to divide the region of interest, $0 \leq x \leq L$, into a selected number of elements. In the one-dimensional case, these elements will be line segments whose endpoints are called nodes. Let $M$ be the number of elements, $N = M + 1$ be the number of nodes, and $l(e)$ ($e = 1, 2, 3, \ldots, M$) be the length of the $e$-th element. Furthermore, let $x_i$ ($i = 1, 2, 3, \ldots, N$) be the coordinate of the $i$-th node. In particular, $x_1 = 0$ and $x_N = L$. The indices $i$ are known as the global node numbers. In addition, we introduce a local numbering system, in which the nodes of the $e$-th element are denoted by $x^{(e)}_1$ and $x^{(e)}_2$, with $x^{(e)}_1 < x^{(e)}_2$. In the present one-dimensional context, the locally and globally numbered coordinates of the nodes are related by

$$x^{(e)}_1 = x_e, \quad x^{(e)}_2 = x_{e+1}; \quad e = 1, 2, 3, \ldots, M$$

Here, on the left-hand sides, the superscript $e$ refers to the element and the subscript (1 or 2) refers to the node's local number. The quantities on the right-hand sides are the globally labeled nodes, and the subscript is a global node number. For example, the two nodes of element 2 are identified locally as $x^{(2)}_1$ and $x^{(2)}_2$ and globally as $x_2$ and $x_3$. The numbering both of elements and of nodes is illustrated in Fig. 15.8, and the relationship between global and local node numbers is conveyed explicitly in the connectivity matrix shown in Table 15.1. Further, with this notation, $l(e) = x_{e+1} - x_e = x^{(e)}_2 - x^{(e)}_1$ and, in general, will vary from element to element.

15.9.2 Selecting Interpolation or Shape Functions

The next step is to select interpolation functions or shape functions that can be used to approximate $\varphi(x)$ within an element. For simplicity, we employ linear functions. Thus, in the $e$-th element, we

---

16 For variety (and to develop the ability to think in both numbering schemes), we here elect to start the numbering of nodes at $i = 1$ rather than at $i = 0$.

17 As noted earlier, higher order polynomials may be used as well, though more nodes would then be necessary. See Exercises 15.10, 15.11, and 15.12.
approximate the unknown function by
\[ \tilde{\varphi}^{(e)} = a^{(e)} + b^{(e)} x \] (15.85)
where \(a^{(e)}\) and \(b^{(e)}\) are constants to be determined. For subsequent convenience, however, it is preferable to express this linear relationship in terms of the values \(\tilde{\varphi}_1^{(e)}\) and \(\tilde{\varphi}_2^{(e)}\) at the end points of the element rather than in terms of the slope \(b^{(e)}\) and intercept \(a^{(e)}\). Since these end points occur locally at \(x_1^{(e)}\) and \(x_2^{(e)}\), we have that
\[ \tilde{\varphi}_1^{(e)} = a^{(e)} + b^{(e)} x_1^{(e)} \]
\[ \tilde{\varphi}_2^{(e)} = a^{(e)} + b^{(e)} x_2^{(e)} \] (15.86)
If we solve Eq. (15.86) for \(a^{(e)}\) and \(b^{(e)}\), substitute the results into Eq. (15.85), and group the terms appropriately, we discover that the approximation in Eq. (15.85) can be written—see Exercise 15.26—in the form
\[ \tilde{\varphi}^{(e)}(x) = \sum_{j=1}^{2} N_j^{(e)}(x) \tilde{\varphi}_j^{(e)} \] (15.87)
where
\[ N_1^{(e)}(x) = \frac{x_2^{(e)} - x}{l^{(e)}} \] (15.88)
\[ N_2^{(e)}(x) = \frac{x - x_1^{(e)}}{l^{(e)}} \] (15.89)
and \(l^{(e)} = x_2^{(e)} - x_1^{(e)}\) is the length of the element. Equations (15.88) and (15.89) define the shape functions for the \(e\)-th element. Graphs of these shape functions are shown in Fig. 15.9. Note that
\[ N_1^{(e)}(x_1^{(e)}) = 1 \quad N_1^{(e)}(x_2^{(e)}) = 0 \]
\[ N_2^{(e)}(x_1^{(e)}) = 0 \quad N_2^{(e)}(x_2^{(e)}) = 1 \] (15.90)
i.e., that one of the shape functions has the value one at the lower end of the element and diminishes linearly to zero at the upper end while the other has the value zero at the lower end and rises linearly to one at the upper end. These properties assure that the sum in Eq. (15.87) has the proper value at each end of the element to which it applies.

### 15.9.3 Formulating the Equations for a Single Element

The next step in the finite element method is to formulate the equations constraining the solution at the nodes defining a single element. We first define the residual \(r(x)\) of Eq. (15.67) as the difference
Figure 15.9: The shape functions for linear interpolation.

between the right-hand side and the left-hand side when the approximate solution \( \tilde{\varphi} \) is substituted into the equation, i.e., by

\[
    r(x) = -\frac{d}{dx} \left( \alpha \frac{d\tilde{\varphi}}{dx} \right) + \beta \tilde{\varphi} - f
\]  

(15.91)

Were \( \tilde{\varphi}(x) \) an exact solution, the residual \( r(x) \) would be identically zero. Since \( \tilde{\varphi} \) is only an approximation to \( \varphi \), however, \( r(x) \) will be nonzero. We define the approximate solution by requiring that \( r(x) \) be as small as possible in some sense. We might, for example, choose to make \( r(x) = 0 \) at a discrete set of points. Still better, we can choose to make an appropriate number of weighted “averages” of \( r(x) \) equal to zero, i.e., we can choose \( \varphi_1^{(e)} \) and \( \varphi_2^{(e)} \) so that

\[
    R_i^{(e)} = \int_{x_1^{(e)}}^{x_2^{(e)}} W_i^{(e)}(x) r(x) \, dx = 0 \quad ; \quad i = 1, 2
\]  

(15.92)

where the \( R_i^{(e)} \) are the weighted residual integrals and the \( W_i^{(e)}(x) \) are as yet unspecified weighting functions for the \( e \)-th element. Note that we need as many weighting functions as we have nodes, since there are that many values of \( \tilde{\varphi}_i^{(e)} \) to determine. Different choices for these weighting functions yield different—though in the end essentially equivalent—methods of solution. In Galerkin’s method, the weighting functions are chosen as the shape functions \( N_i^{(e)}(x) \) used for the expansion of \( \tilde{\varphi} \) in Eq. (15.87).\(^{18}\) With that choice, the weighted residual integral for the \( e \)-th element is given by

\[
    R_i^{(e)} = \int_{x_1^{(e)}}^{x_2^{(e)}} N_i^{(e)} \left[ -\frac{d}{dx} \left( \alpha(x) \frac{d\tilde{\varphi}_i^{(e)}}{dx} \right) + \beta(x) \tilde{\varphi}_i^{(e)} \right] \, dx - \int_{x_1^{(e)}}^{x_2^{(e)}} N_i^{(e)} f(x) \, dx \quad i = 1, 2
\]  

(15.93)

where Eq. (15.91) was substituted into Eq. (15.92) for \( r(x) \). If the first term is integrated by parts, we find that\(^{19}\)

\[
    R_i^{(e)} = \int_{x_1^{(e)}}^{x_2^{(e)}} \left( \alpha(x) \frac{dN_i^{(e)}}{dx} \frac{d\tilde{\varphi}_i^{(e)}}{dx} + \beta(x) N_i^{(e)} \tilde{\varphi}_i^{(e)} \right) \, dx
\]

\(^{18}\)Alternatively, we could use the Ritz method to formulate the system of equations. The Ritz method is a variational method in which the boundary value problem is formulated in terms of a variational expression or functional. The minimum of this functional corresponds to the governing differential equation under the given boundary conditions. To obtain the approximate solution, the functional is minimized with respect to its variables. In this approach, the choice of the weighting functions, which seemed quite arbitrary in the Galerkin approach, is embedded naturally in the development. In all cases where a variational formulation exists, the Galerkin and Ritz methods are equivalent. The Galerkin method is, however, applicable even in cases for which a variational formulation cannot be found.

\(^{19}\)Recall that \( \int u \, dv = uv - \int v \, du \) where for this case \( u = N_i^{(e)} \) and \( dv = (d/dx) (\alpha^{(e)} \frac{d\tilde{\varphi}_i^{(e)}}{dx}) \, dx \).
Now if Eq. (15.87) is substituted into Eq. (15.94), we find that

\[
R_i^{(e)} = \sum_{j=1}^{2} \int_{x_1^{(e)}}^{x_2^{(e)}} \left( \alpha(x) \frac{dN_i^{(e)}}{dx} - \frac{d\phi^{(e)}}{dx} \right) dx
- \int_{x_1^{(e)}}^{x_2^{(e)}} N_i^{(e)} f(x) dx - \left( \alpha N_i^{(e)} \frac{d\phi^{(e)}}{dx} \right) \bigg|_{x=x_1^{(e)}}^{x=x_2^{(e)}}; \quad i = 1, 2. \tag{15.95}
\]

These equations for the weighted residual integrals can also be expressed in matrix form as

\[
\begin{pmatrix}
R_1^{(e)} \\
R_2^{(e)}
\end{pmatrix} =
\begin{bmatrix}
K_{11}^{(e)} & K_{12}^{(e)} \\
K_{21}^{(e)} & K_{22}^{(e)}
\end{bmatrix}
\begin{pmatrix}
\alpha N_i^{(e)} \\
\phi^{(e)}
\end{pmatrix}
- \begin{pmatrix}
b_1^{(e)} \\
b_2^{(e)}
\end{pmatrix}
- \begin{pmatrix}
g_1^{(e)} \\
g_2^{(e)}
\end{pmatrix} \tag{15.96}
\]

or more compactly as

\[
\begin{pmatrix}
R_1^{(e)} \\
R_2^{(e)}
\end{pmatrix} = [K^{(e)}] \{ \alpha N_i^{(e)} \} - \{ b^{(e)} \} - \{ g^{(e)} \} \tag{15.97}
\]

where

\[
K_{ij}^{(e)} = \int_{x_1^{(e)}}^{x_2^{(e)}} \left( \alpha(x) \frac{dN_i^{(e)}}{dx} \frac{dN_j^{(e)}}{dx} + \beta(x) N_i^{(e)} N_j^{(e)} \right) dx \tag{15.98}
\]

\[
b_i^{(e)} = \int_{x_1^{(e)}}^{x_2^{(e)}} N_i^{(e)} f(x) dx \tag{15.99}
\]

and

\[
g_i^{(e)} = \left( \alpha N_i^{(e)} \frac{d\phi^{(e)}}{dx} \right) \bigg|_{x=x_1^{(e)}}^{x=x_2^{(e)}} = \left( \alpha N_i^{(e)} \frac{d\phi^{(e)}}{dx} \right) \bigg|_{x=x_1^{(e)}}^{x=x_2^{(e)}} \tag{15.100}
\]

Note that the matrix \(K\) is necessarily symmetric in \(i\) and \(j\).

In general, the elements will be short, and the functions \(\alpha(x)\) and \(\beta(x)\) will be slowly varying functions of \(x\). Over the range of integration appearing in Eq. (15.98), these functions can often be treated as constants, though with different values for each element. Under those circumstances, Eq. (15.98) can be evaluated analytically by taking the derivatives of the shape functions, substituting those derivatives and the shape functions themselves into Eq. (15.98), and integrating over the element. The result is a \(2 \times 2\) matrix whose elements are\(^{20}\)

\[
K_{11}^{(e)} = K_{22}^{(e)} = \frac{\alpha^{(e)}}{l^{(e)}} + \frac{\beta^{(e)} l^{(e)}}{3} \tag{15.101}
\]

\[
K_{12}^{(e)} = K_{21}^{(e)} = -\frac{\alpha^{(e)}}{l^{(e)}} + \frac{\beta^{(e)} l^{(e)}}{6} \tag{15.102}
\]

where, in these equations, \(\alpha^{(e)}\) and \(\beta^{(e)}\) stand for approximate (constant) values of \(\alpha(x)\) and \(\beta(x)\) appropriate to the \(e\)-th element, and \(l^{(e)} = x_2^{(e)} - x_1^{(e)}\) is the length of the \(e\)-th element. Similarly, if \(f^{(e)}\) is the (approximate) constant value of \(f\) within the \(e\)-th element, Eq. (15.99) can be evaluated to give

\[
b_1^{(e)} = b_2^{(e)} = \frac{f^{(e)} l^{(e)}}{2} \tag{15.103}
\]

\(^{20}\)The integrals that appear here can, of course, be evaluated manually. An alternative evaluation using a symbol manipulating program is laid out in Appendix 15.B.
If, finally, we replace \( N_1^{(e)} \) and \( N_2^{(e)} \) in Eq. (15.100) with the shape functions in Eqs. (15.88) and (15.89), we have that

\[
g_{1}^{(e)} = -\alpha(x_1^{(e)}) \left. \frac{d\bar{\varphi}_{1}^{(e)}}{dx} \right|_{x_1^{(e)}}
\]

\[
g_{2}^{(e)} = \alpha(x_2^{(e)}) \left. \frac{d\bar{\varphi}_{2}^{(e)}}{dx} \right|_{x_2^{(e)}}
\]

With \([K]\), \(\{b\}\), and \(\{g\}\) given by Eqs. (15.101)–(15.104), the *elemental equation* obtained by requiring \(R_{1}^{(e)}\) as given by Eq. (15.96) to be zero for all \(i\) is

\[
\begin{bmatrix}
K_{11}^{(e)} & K_{12}^{(e)} \\
K_{21}^{(e)} & K_{22}^{(e)}
\end{bmatrix}
\begin{bmatrix}
\bar{\varphi}_{1}^{(e)} \\
\bar{\varphi}_{2}^{(e)}
\end{bmatrix}
= 
\begin{bmatrix}
b_{1}^{(e)} \\
b_{2}^{(e)}
\end{bmatrix} + 
\begin{bmatrix}
g_{1}^{(e)} \\
g_{2}^{(e)}
\end{bmatrix}
\tag{15.105}
\]

### 15.9.4 Assembling the System of Equations

The next step is to assemble the elemental equations for *each* element into a single (large) set describing all elements. The process of assembly is best illustrated with an example. Suppose we divide the interval \(0 \leq x \leq L\) into \(M = 3\) elements with \(N = M + 1 = 4\) nodes. The elemental equation for the first element is given by Eq. (15.105) with the superscript \((e)\) set to one, i.e., by

\[
\begin{bmatrix}
K_{11}^{(1)} & K_{12}^{(1)} \\
K_{21}^{(1)} & K_{22}^{(1)}
\end{bmatrix}
\begin{bmatrix}
\bar{\varphi}_{1}^{(1)} \\
\bar{\varphi}_{2}^{(1)}
\end{bmatrix}
= 
\begin{bmatrix}
b_{1}^{(1)} \\
b_{2}^{(1)}
\end{bmatrix} + 
\begin{bmatrix}
g_{1}^{(1)} \\
g_{2}^{(1)}
\end{bmatrix}
\tag{15.106}
\]

which is equivalent to the two equations

\[
K_{11}^{(1)} \bar{\varphi}_{1}^{(1)} + K_{12}^{(1)} \bar{\varphi}_{2}^{(1)} = b_{1}^{(1)} + g_{1}^{(1)}
\]

\[
K_{21}^{(1)} \bar{\varphi}_{1}^{(1)} + K_{22}^{(1)} \bar{\varphi}_{2}^{(1)} = b_{2}^{(1)} + g_{2}^{(1)}
\]

Note that in Eqs (15.106) and (15.107) we have employed the *local* numbering system. For the other two elements—also with local node numbers—we have similarly that

\[
K_{11}^{(2)} \bar{\varphi}_{1}^{(2)} + K_{12}^{(2)} \bar{\varphi}_{2}^{(2)} = b_{1}^{(2)} + g_{1}^{(2)}
\]

\[
K_{21}^{(2)} \bar{\varphi}_{1}^{(2)} + K_{22}^{(2)} \bar{\varphi}_{2}^{(2)} = b_{2}^{(2)} + g_{2}^{(2)}
\]

and

\[
K_{11}^{(3)} \bar{\varphi}_{1}^{(3)} + K_{12}^{(3)} \bar{\varphi}_{2}^{(3)} = b_{1}^{(3)} + g_{1}^{(3)}
\]

\[
K_{21}^{(3)} \bar{\varphi}_{1}^{(3)} + K_{22}^{(3)} \bar{\varphi}_{2}^{(3)} = b_{2}^{(3)} + g_{2}^{(3)}
\]

From the relationship between the local and global numbering systems as defined in Eq. (15.84), however, it is clear that \(x_2^{(1)} = x_2^{(2)} = x_2^{(3)} = x_3\). Imposing the condition of continuity at nodes 2 and 3, we conclude then that \(\bar{\varphi}_{2}^{(1)} = \bar{\varphi}_{1}^{(2)}\) and \(\bar{\varphi}_{2}^{(2)} = \bar{\varphi}_{1}^{(3)}\). Now if we once again invoke the relationship in Eq. (15.84) and recognize the correspondences

\[
\bar{\varphi}_{1}^{(1)} \rightarrow \bar{\varphi}_{1}, \quad \bar{\varphi}_{2}^{(1)} \rightarrow \bar{\varphi}_{2}, \quad \bar{\varphi}_{2}^{(2)} = \bar{\varphi}_{1}^{(3)} \rightarrow \bar{\varphi}_{3}, \quad \text{and} \quad \bar{\varphi}_{2}^{(3)} \rightarrow \bar{\varphi}_{4}
\]
linking the (approximate) values at *locally* numbered nodes to the (approximate) values \(\tilde{\varphi}_1, \tilde{\varphi}_2, \tilde{\varphi}_3,\) and \(\tilde{\varphi}_4\) at the *globally* numbered nodes, Eqs. (15.107)–(15.109) become

\[
\begin{align*}
K_{11}^{(1)} \tilde{\varphi}_1 + K_{12}^{(1)} \tilde{\varphi}_2 &= b_1^{(1)} + g_1^{(1)} \\
K_{21}^{(1)} \tilde{\varphi}_1 + K_{22}^{(1)} \tilde{\varphi}_2 &= b_2^{(1)} + g_2^{(1)} \\
K_{11}^{(2)} \tilde{\varphi}_2 + K_{12}^{(2)} \tilde{\varphi}_3 &= b_1^{(2)} + g_1^{(2)} \\
K_{21}^{(2)} \tilde{\varphi}_2 + K_{22}^{(2)} \tilde{\varphi}_3 &= b_2^{(2)} + g_2^{(2)} \\
K_{11}^{(3)} \tilde{\varphi}_3 + K_{12}^{(3)} \tilde{\varphi}_4 &= b_1^{(3)} + g_1^{(3)} \\
K_{21}^{(3)} \tilde{\varphi}_3 + K_{22}^{(3)} \tilde{\varphi}_4 &= b_2^{(3)} + g_2^{(3)}.
\end{align*}
\]  

(15.110)

Thus, we have six equations but only four unknowns. Some of these equations must be redundant. If we choose carefully, we might ignore two of them. Alternatively, we can reduce the number of equations to four by replacing the second and third equations with their sum and, similarly, replacing the fourth and fifth equations with their sum to find that Eq. (15.110) becomes

\[
\begin{bmatrix}
K_{11}^{(1)} & K_{12}^{(1)} & 0 & 0 \\
K_{21}^{(1)} & K_{22}^{(1)} + K_{11}^{(2)} & K_{12}^{(2)} & 0 \\
0 & K_{21}^{(2)} & K_{22}^{(2)} + K_{11}^{(3)} & K_{12}^{(3)} \\
0 & 0 & K_{21}^{(3)} & K_{22}^{(3)}
\end{bmatrix}
\begin{bmatrix}
\tilde{\varphi}_1 \\
\tilde{\varphi}_2 \\
\tilde{\varphi}_3 \\
\tilde{\varphi}_4
\end{bmatrix}
= \begin{bmatrix}
b_1^{(1)} \\
b_2^{(1)} + b_1^{(2)} \\
b_2^{(2)} + b_1^{(3)} \\
b_2^{(3)} + g_2^{(3)}
\end{bmatrix} + \begin{bmatrix}
g_1^{(1)} \\
g_2^{(1)} + g_2^{(2)} \\
g_2^{(2)} + g_2^{(3)} \\
g_2^{(3)}
\end{bmatrix}.
\]

(15.111)

or more compactly

\[
[K] \{\tilde{\varphi}\} = \{b\} + \{g\}.
\]  

(15.112)

In Eq. (15.112), \([K]\) is the assembled *stiffness* matrix for the three-element problem. The extension to more than three elements is now evident. Note that, no matter how many elements we have in a one-dimensional problem divided into elements with two nodes, the matrix \([K]\) will always be tridiagonal.

We complete this step in the process by working out the elements in the assembled equation. From Eqs. (15.101), (15.102), and (15.111), we conclude that the nonzero elements of \([K]\) can be written as

\[
K_{ii} = K_{11}^{(i)} - \frac{\alpha^{(i)} l^{(i)}}{3} \quad + \beta^{(i)} \frac{l^{(i)}}{3} \\
K_{NN} = K_{22}^{(M)} - \frac{\alpha^{(M)} l^{(M)}}{3} \quad + \beta^{(M)} \frac{l^{(M)}}{3} \\
K_{i+1,i} = K_{i,i+1} = K_{12}^{(i)} - \frac{\alpha^{(i)} l^{(i)}}{6} \quad + \beta^{(i)} \frac{l^{(i)}}{6}
\]

(15.113)

where (by way of reminder) \(N = M + 1\) is the number of nodes and \(M\) is the number of elements. Similarly, using Eqs. (15.103) and (15.111), we can write the elements of \(\{b\}\) in the form

\[
b_1 = b_1^{(1)} = \frac{f^{(1)} l^{(1)}}{2}
\]
\[ b_i = b_2^{(i-1)} + b_1^{(i)} = f^{(i-1)} \frac{l^{(i-1)}}{2} + f^{(i)} \frac{l^{(i)}}{2} \quad i = 2, 3, 4, \ldots, N - 1 \]

\[ b_N = b_1^{(M)} = f^{(M)} \frac{l^{(M)}}{2} \]

Finally, we need to evaluate the elements of \( \{g\} \). Note that all but the first and last entries in \( \{g\} \) can be written as

\[ g_i = g_2^{(i-1)} + g_1^{(i)} \]

If Eq. (15.104) is substituted into Eq. (15.115), we have

\[ g_i = \alpha \left( \frac{d\hat{\varphi}}{dx} \right) \bigg|_{x=x_i} - \alpha \left( \frac{d\hat{\varphi}}{dx} \right) \bigg|_{x=x_i} \]

However, \( x_2^{(i-1)} = x_1^{(i)} = x_i \) [see Eq. (15.84)] and \( \alpha (d\hat{\varphi}/dx) \) must be continuous at \( x_i \). As a result, the right-hand side of Eq. (15.116) is zero for \( i = 2, 3 \) and we are left with the vector

\[ \{ g \} = \begin{bmatrix} -\alpha(x_1) \left( \frac{d\hat{\varphi}}{dx} \right) \bigg|_{x=x_1} & 0 & 0 & \alpha(x_4) \left( \frac{d\hat{\varphi}}{dx} \right) \bigg|_{x=x_4} \end{bmatrix} \]

In general for a problem with \( M \) elements (\( N \) nodes), we would find that

\[ g_1 = -\alpha(x_1) \left( \frac{d\hat{\varphi}}{dx} \right) \bigg|_{x=x_1=0} \]

\[ g_i = 0 \quad i = 2, 3, 4, \ldots, N - 1 \]

\[ g_N = \alpha(x_N) \left( \frac{d\hat{\varphi}}{dx} \right) \bigg|_{x=x_N=L} \]

In consequence of the considerations to this point, our system of equations for \( \{ \hat{\varphi} \} \)—the (approximate) values at the nodes—now assumes the form \(^{21}\)

\[
\begin{bmatrix}
K_{11} & K_{12} & K_{13} & K_{14} \\
K_{21} & K_{22} & K_{23} & K_{24} \\
K_{31} & K_{32} & K_{33} & K_{34} \\
K_{41} & K_{42} & K_{43} & K_{44}
\end{bmatrix}
\begin{bmatrix}
\hat{\varphi}_1 \\
\hat{\varphi}_2 \\
\hat{\varphi}_3 \\
\hat{\varphi}_4
\end{bmatrix}
= \begin{bmatrix}
b_1 \\
b_2 \\
b_3 \\
b_4
\end{bmatrix}
+ \begin{bmatrix}
g_1 \\
0 \\
0 \\
g_4
\end{bmatrix}
\]

Similar expressions can be readily written for cases in which there are more nodes than four.

### 15.9.5 Incorporating the Boundary Conditions

Before we can solve Eq. (15.119) for \( \{ \hat{\varphi} \} \), we need to incorporate the boundary conditions. First consider the boundary condition of the third kind given in Eq. (15.69). If we replace \( \varphi \) by its approximation \( \hat{\varphi} \), solve for \( \alpha (d\hat{\varphi}/dx) \) at \( x = L \), and substitute this into Eq. (15.118), we have that

\[ g_N = \alpha \left( \frac{d\hat{\varphi}}{dx} \right) \bigg|_{x=L} = q - \gamma \hat{\varphi}_N \]

\(^{21}\)We elect to introduce a symbol for every element in \( K \) even though many of those elements are known in advance to be zero.
or specifically for our case that
\[ g_4 = \alpha \frac{d\tilde{\phi}}{dx} \bigg|_{x=L} = q - \gamma \tilde{\phi}_4 \]  
\tag{15.121}

Thus, the last equation in our system becomes
\[ K_{41} \tilde{\phi}_1 + K_{42} \tilde{\phi}_2 + K_{43} \tilde{\phi}_3 + K_{44} \tilde{\phi}_4 = b_4 + q - \gamma \tilde{\phi}_4 \]  
\tag{15.122}

which we can recast as
\[ K_{41} \tilde{\phi}_1 + K_{42} \tilde{\phi}_2 + K_{43} \tilde{\phi}_3 + (K_{44} + \gamma) \tilde{\phi}_4 = b_4 + q \]  
\tag{15.123}

With this rewriting of the fourth equation, our system of equations becomes
\[
\begin{bmatrix}
K_{11} & K_{12} & K_{13} & K_{14} \\
K_{21} & K_{22} & K_{23} & K_{24} \\
K_{31} & K_{32} & K_{33} & K_{34} \\
K_{41} & K_{42} & K_{43} & K_{44} + \gamma
\end{bmatrix} \begin{bmatrix}
\tilde{\phi}_1 \\
\tilde{\phi}_2 \\
\tilde{\phi}_3 \\
\tilde{\phi}_4
\end{bmatrix} = \begin{bmatrix}
b_1 \\
b_2 \\
b_3 \\
b_4 + q
\end{bmatrix} + \begin{bmatrix}
g_1 \\
0 \\
0 \\
0
\end{bmatrix}
\]  
\tag{15.124}

In general for a domain with \( N \) nodes, we have
\[ K_{NN} \rightarrow K_{NN} + \gamma = \frac{\alpha^{(M)}}{l^{(M)}} + \frac{\beta^{(M)}}{3} + \gamma \]  
\tag{15.125}

and hence the \( N \)th element of \( \{g\} \) is absorbed in \([K]\) and \( \{b\} \). This operation can always be performed for mixed (and Neumann) boundary conditions.

Imposing the Dirichlet boundary condition of Eq. (15.68) is simpler. Recall that for \( M \) elements, we have \( M+1 \) unknowns \( \tilde{\phi}_i \) \((i = 1, 2, 3, \ldots, N)\) and \( M+1 \) equations. However, the Dirichlet boundary condition given in Eq. (15.68) specifies the value of one of these unknowns, specifically \( \tilde{\phi}_1 \). Thus we actually have \( M+2 \) equations that need to be simultaneously satisfied. But since we only need as many equations as unknowns, we can replace the first equation of Eq. (15.124) with Eq. (15.68). As a result, \( g_1 \) no longer plays a role in the system of equations, and the entire \( g \) vector has been absorbed.\(^{22}\) The new system of equations for our three-element example is therefore
\[
\begin{bmatrix}
1 & 0 & 0 & 0 \\
K_{21} & K_{22} & K_{23} & K_{24} \\
K_{31} & K_{32} & K_{33} & K_{34} \\
K_{41} & K_{42} & K_{43} & K_{44}
\end{bmatrix} \begin{bmatrix}
\tilde{\phi}_1 \\
\tilde{\phi}_2 \\
\tilde{\phi}_3 \\
\tilde{\phi}_4
\end{bmatrix} = \begin{bmatrix}
p \\
b_2 \\
b_3 \\
b_4
\end{bmatrix}
\]  
\tag{15.127}

where \( K_{44} \) and \( b_4 \) have been modified according to Eqs. (15.125) and (15.126).

The symmetry of the coefficient matrix is easily restored. Consider, for example, the second equation in Eq. (15.127), namely
\[ K_{21} \tilde{\phi}_1 + K_{22} \tilde{\phi}_2 + K_{23} \tilde{\phi}_3 + K_{24} \tilde{\phi}_4 = b_2 \]  
\tag{15.128}

Since \( \tilde{\phi}_1 = p \), however, this equation can be recast as
\[ K_{22} \tilde{\phi}_2 + K_{23} \tilde{\phi}_3 + K_{24} \tilde{\phi}_4 = b_2 - K_{21} p \]  
\tag{15.129}

\(^{22}\)Actually the original first equation \( K_{11} \tilde{\phi}_1 + K_{12} \tilde{\phi}_2 + K_{13} \tilde{\phi}_3 + K_{14} \tilde{\phi}_4 = b_1 + g_1 \) is merely reinterpreted as an equation determining \( g_1 \) after the solution \( \{ \tilde{\phi} \} \) has been found.
A similar recasting of the other equations leads finally to

\[
\begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & K_{22} & K_{23} & K_{24} \\
0 & K_{32} & K_{33} & K_{34} \\
0 & K_{42} & K_{43} & K_{44}
\end{bmatrix}
\begin{bmatrix}
\tilde{\varphi}_1 \\
\tilde{\varphi}_2 \\
\tilde{\varphi}_3 \\
\tilde{\varphi}_4
\end{bmatrix}
= \begin{bmatrix}
p \\
b_2 - K_{21}p \\
b_3 - K_{31}p \\
b_4 - K_{41}p
\end{bmatrix}
\] (15.130)

and once again we have a symmetric system. Further, since all elements in the coefficient matrix and all elements in the vector of inhomogeneities are now known, these four equations uniquely determine the vector \{\tilde{\varphi}\} of unknowns.\textsuperscript{23}

We can carry our analysis no further without selecting a particular programming language in which to implement explicit coding. That task is undertaken in the next several sections.

15.12 Using OCTAVE to Solve 1D PDEs via an FEM

15.12.1 A General Coding

The final step is to solve Eq. (15.130) for the unknowns \(\tilde{\varphi}_0, \tilde{\varphi}_1, \ldots, \tilde{\varphi}_M\). A program to construct Eq. (15.130) and then solve the resulting system for \(\tilde{\varphi}\) at each node would begin by asking the user to input the values of the various parameters involved in the problem. In all cases, the parameters \(p, \gamma,\) and \(q\), which relate to the boundary conditions, are constants. In general, the quantities \(\alpha, \alpha',\) and \(\beta\), which relate to the physical system involved, and \(f\), which represents a source or excitation function will depend on \(x\). Further, \(l(e)\), which reflects the particular discretization adopted, will in general depend on the element. For simplicity in illustration, however, we will suppose these five quantities to be constants as well.\textsuperscript{24} Indeed, if \(\alpha\) is constant, then \(\alpha'\) is zero, and we will simply leave out terms multiplied by \(\alpha'\) when we construct the equations.

The first segment of the program we wish to write will, then, request input of all the constants needed in the remainder of the program. Appropriate OCTAVE statements (BLOCK 1), including an evaluation of the length of the rod and the coordinates of the nodes, are

\[
\begin{align*}
M &= \text{input}(\ 'Enter number of segments (M): ' ); \\
alpha &= \text{input}(\ 'Enter alpha: ' ); \\
beta &= \text{input}(\ 'Enter beta: ' ); \\
f &= \text{input}(\ 'Enter f: ' ); \\
l &= \text{input}(\ 'Enter l: ' ); \\
p &= \text{input}(\ 'Enter p: ' ); \\
gamma &= \text{input}(\ 'Enter gamma: ' ); \\
q &= \text{input}(\ 'Enter q: ' ); \\
length &= l*\text{M} ; \\
x &= l * \[ 0 : \text{M} ];
\end{align*}
\]

The next step is to assemble \([K]\) as given in Eqs. (15.111) and (15.113). To this end, we define \(K\) as an \((M+1) \times (M+1)\) floating array with all elements set to zero. Then, we assign values to the diagonal and non-zero off-diagonal elements in accordance with Eq. (15.113) with the statements\textsuperscript{25} (BLOCK 2)

\textsuperscript{23}We assume, of course, that the coefficient matrix is not singular.

\textsuperscript{24}If \(\alpha, \alpha', \beta, f,\) and \(l(e)\) are not constant, each would have to be represented by an appropriately sized vector, and values would have to be given for all elements in these vectors.

\textsuperscript{25}Remember that indices in OCTAVE start at one rather than zero, so all OCTAVE indices coincide with the corresponding subscripts on \(K\).
K = zeros(M+1,M+1); % Create (M+1)x(M+1) array of zeros
S = alpha/l + beta*l/3.0; % Evaluate common quantities
S2 = 2.0*S;
T = -alpha/l + beta*l/6.0;
K(1,1) = S; % Set diagonal elements of K
for i = 2:M K(i,i) = S2; end
K(M+1,M+1) = S;
for i = 1:M K(i+1,i) = T; end % Set elements above and below main diagonal of K
for i = 1:M K(i,i+1) = T; end % main diagonal of K

To complete the assembly, we create the vector \( \{b\} \) as given by Eq. (15.114) in a similar fashion, invoking the statements (BLOCK 3)

b = zeros(M+1,1); % Create M+1 element vector of zeros
U = f*l/2.0; % Evaluate common quantities
U2 = 2.0*U;
b(1) = U; % Set elements of b
for i = 2:M b(i) = U2; end
b(M+1) = U;

Finally, to produce the system of equations given in Eq. (15.130), we must incorporate the boundary conditions and thus modify \( K \) and \( b \). First, we impose the boundary condition of the third kind (mixed conditions) through Eqs. (15.125) and (15.126) with the statements (BLOCK 4)

K(M+1,M+1) = K(M+1,M+1) + gamma;
b(M+1) = b(M+1) + q;

Then we impose the Dirichlet condition as described in Section 15.9.5 with the statements (BLOCK 5)

K(1,1) = 1.0;
b(1) = p;
for i = 2:M+1 K(i,1) = 0.0; end
for i = 2:M+1 b(i) = b(i) - K(1,i)*p; end
for j = 2:M+1 K(1,j) = 0.0; end

As in Eq. (15.130), \( K_{11} \) is set equal to \( p \). Then, we assign the value zero to \( K_{ij} \) for \( j = 2, 3, 4, \ldots, M \). Finally, the last two lines multiply all but the first entry in the first column of \( K \) by \( p \), subtract each of those values from the corresponding element of \( b \) and then—note the order of operations—set all but the first element in the first column of \( K \) to zero. As a result, we now have in \( K \) and \( b \) the coefficient matrix and vector of inhomogenieties as in Eq. (15.130), a symmetric system of equations ready to be solved.

To solve this system of equations, which is \( [K]\{\tilde{\phi}\} = \{b\} \), we invoke the simple OCTAVE statement (BLOCK 6)

\[
\tilde{\phi} = K\backslash b;
\]

Upon return to the OCTAVE command prompt, \( \tilde{\phi} \) contains the (approximate) solution to the boundary value problem defined in Section 15.1.7 and \( x \) contains the coordinates of the nodes along the rod.
A more fully commented command file containing these statements is named `fem1d.m`, is listed in Appendix 15.C.3, and can be copied from the directory `$HEAD/octave`.\(^{26}\)

### 15.12.2 An Example: Simple Harmonic Motion

We now illustrate the application of `fem1d` to a simplification of Eq. (15.67), specifically the equation

\[
md^2\varphi/dt^2 + k\varphi = 0 \quad ; \quad 0 \leq t \leq T
\]  

(15.131)

for the simple harmonic motion of a mass \(m\) attached to a spring of stiffness \(k\). To match this situation to that discussed in Section 15.1.7, we must interpret \(\alpha\) as a constant equal to \(-m\), \(\beta\) as a constant equal to \(k\), \(L\) as the time \(T\) at the end of the interval of interest, and the independent variable \(x\) as the independent variable \(t\). Further, we must set \(f = 0\). The dependent variable \(\varphi\) gives the displacement of the oscillator from its equilibrium position. Basically, we are interested in the position over a range of times \(0 \leq t \leq T\). In a consistent set of units, we will take

\[
m = -\alpha = 4.0 \text{ kg} \quad ; \quad k = \beta = 3.0 \text{ N/m}
\]

To complete the definition of the problem, we need to specify the boundary conditions.\(^{27}\) Suppose we seek a solution for which \(\varphi(0) = 0\) m and \(d\varphi(t)/dt|_{t=T} = 1.0\) m/s, i.e., we specify the position at \(t = 0\) s and the velocity at \(t = T\). To effect these conditions, we need to assign the values

\[
p = 0.0 \text{ m} \quad ; \quad \gamma = 0.0 \text{ kg/s} \quad ; \quad q = -4.0 \text{ kg-m/s}
\]

the first of which reduces Eq. (15.68) to \(\varphi(0) = 0\) m and the rest of which (with \(\alpha = -4.0\) kg) reduce Eq. (15.69) to \(d\varphi(t)/dt|_{t=T} = 1.0\) m/s. In this example, we will seek a solution over the interval \(0 \leq t \leq 10\) s, so \(L \rightarrow T = 10.0\) s. In physical terms, we start the oscillator at its equilibrium position with an unspecified initial velocity such that, at \(t = 10.0\) s, its velocity will be 1.0 m/s.

With these choices, we are now ready to execute the command file and solve the problem. The statement

```
fem1d
```

will execute the statements in the file, one at a time. The first several statements will request input, to which we will respond with the values\(^{28}\)

```
Enter number of segments (M): 20
Enter alpha: -4.0
Enter beta: 3.0
Enter f: 0.0
Enter l: 0.5
Enter p: 0.0
Enter gamma: 0.0
Enter q: -4.0
```

Once we have entered these parameters, the remaining statements will construct all necessary matrices and vectors and then generate the solution, storing it in an array named `phi`. We will have

\(^{26}\)At some sites, the file may also be located in the OCTAVE directory structure such that OCTAVE can find it when it is identified only by its name without a prepended path.

\(^{27}\)Note that we are using a method that requires one item of information at each end of the region of interest—a boundary value problem. More often in problems in motion, one has an initial value problem in which one specifies two items at one end of the region of interest, say an initial position and an initial velocity.

\(^{28}\)Remember that we have written the command file `fem1d.pro` to expect \(l = L/M\).
no further interaction with the program and, presently, the OCTAVE prompt will return. At that point, all variables to which values have been assigned in the execution of the command file will be accessible at OCTAVE’s main prompt. In particular, we could plot the solution by invoking the statements

```octave
plot( x, phi, 'color', 'black', 'linewidth', 4);
title ( 'M = 20', 'fontsize', 16 );
grid on
```

The resulting graph is shown in the upper left panel of Fig. 15.10.

To assess the accuracy of this solution, we repeat the process with an increasing number of divisions of the interval in which the solution is sought. To save the solution just generated and generate others for $M = 50$ and $M = 100$, we execute the statements

```octave
x020 = x; phi020 = phi;
fem1d
Enter number of segments (M): 50
   . (rest same as for M=20, except l = 10/50 = 0.2)
x050 = x; phi050 = phi;
fem1d
Enter number of segments (M): 100
   . (rest same as for M=20, except l = 10/100 = 0.1)
x100 = x; phi100 = phi;
```

Then, to plot all solutions in a single display, we execute the statements

```octave
subplot(2,2,1)
plot( x020, phi020, 'Color', 'black', 'LineWidth', 4);
title ( 'M = 20', 'fontsize', 16 );
grid on
subplot(2,2,2)
plot( x050, phi050, 'Color', 'black', 'LineWidth', 4);
title ( 'M = 50', 'fontsize', 16 );
grid on
subplot(2,2,3)
plot( x100, phi100, 'Color', 'black', 'LineWidth', 4);
title ( 'M = 100', 'fontsize', 16 );
grid on
subplot(2,2,4)
plot( x020, phi020, 'Color', 'black', 'LineWidth', 4);
hold on
plot( x050, phi050, 'Color', 'black', 'LineWidth', 4);
plot( x100, phi100, 'Color', 'black', 'LineWidth', 4);
title ( 'M = 20, 50, 100', 'fontsize', 16 );
grid on
hold off
```

All of these graphs are shown in Fig. 15.10. Further, as revealed most directly in the lower right panel, even the solution with only 20 segments lies quite close to the solutions for 50 and 100 segments. We conclude that the solution we have obtained is accurate, at least to the resolution of the graph paper.
As a further test on the accuracy of the solution, which we expect to improve as the number of segments increases, we compare the solution for $M = 50$ with that for $M = 100$ with the statements\footnote{This loop stops at $i = 50$ because going to $i = 51$ ultimately seeks $\phi(102)$, which is non-existent.}

\begin{verbatim}
for i=1:50 diff(i) = phi100(2*i) - phi050(i); end
[max(diff), min(diff)]
ans = 0.14437 -0.14899
[max(phi100), min(phi100)]
ans = 1.6035 -1.6045
\end{verbatim}

which suggests that the solution for $M = 50$ is accurate to an absolute error of about $\pm 0.15$ and that the solution falls in the range $-1.6 \leq \phi \leq 1.6$. If we generate a solution for $M = 200$, its difference from the solution for $M = 100$ ranges from $-0.0718$ to $0.0706$, so the solution for $M = 100$ evidently has an absolute error of about $\pm 0.07$.

The above tests of accuracy have, of course, assessed only discretization error. To test for roundoff error, we would have to increase the number of nodes even further, looking for the point at which the solution begins to depart from that to which it appears at the moment to be converging. Since we have in this example used a direct method for finding the solution to the system of linear equations, we need not be concerned about convergence error.

Finally, using the solution for $M = 100$, we might determine the initial velocity and check the final velocity with the statements

\begin{verbatim}
(phi100(2)-phi100(1))/ ... Display initial velocity
(x100(2)-x100(1))
-1.3875
(phi100(101)-phi100(100))/ ... Display final velocity
(x100(101)-x100(100))
0.95704
\end{verbatim}
Despite the graphical agreement of the solutions for $M = 50$ and $M = 100$, the final velocity for $M = 100$ doesn’t quite match the prescribed boundary value of 1.0 m/s. For comparison, had we used $M = 200$, we would have found the initial and final velocities to be $-1.3861$ m/s and $0.97890$ m/s, respectively; had we used $M = 500$, we would have found those velocities to be $-1.3857$ and $0.99164$ m/s. Apparently, with increasing $M$, the calculated final velocity converges on the prescribed boundary value.

15.16 Finite Difference Methods (FDMs) in Two Dimensions

We can easily extend the one-dimensional FDM already described to apply to problems involving two independent variables. We must, however, treat initial value problems, such as those involving the wave and diffusion equations, differently from boundary value problems, such as those involving the Laplace equation.

15.16.1 The Wave Equation

As laid out in Section 15.1.1, the wave equation for waves in one dimension involves two independent variables, a spatial coordinate $x$ locating a point in the one-dimensional medium and a time coordinate $t$. The displacement $u(x,t)$ of the medium satisfies Eq. (15.13),

$$\frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}$$  \hspace{1cm} (15.132)

which is to be solved subject both to the initial conditions

$$u(x,0) = f(x) \quad ; \quad \frac{\partial u}{\partial t}(x,0) = g(x)$$  \hspace{1cm} (15.133)

that specify how the motion of the string is initiated and to appropriate boundary conditions that specify how the medium is “fastened” at its two ends, say at $x = 0$ and $x = L$. To illustrate two distinct types of boundary conditions, we suppose here that the displacement of the medium must be zero for all time at $x = 0$ and the derivative of that displacement must be zero for all time at $x = L$, i.e., that

$$u(0,t) = 0 \quad ; \quad \frac{\partial u}{\partial x}(L,t) = 0$$  \hspace{1cm} (15.134)

For a string, for which $u(x,t)$ represents a displacement, these conditions stipulate physically that the string is tied down at $x = 0$ and free at $x = L$; for an air column, for which $u(x,t)$ also represents a displacement, these conditions stipulate that the pipe containing the column is closed at $x = 0$ (where there will therefore be a displacement node and a pressure antinode) and open at $x = L$ (where there will be a displacement antinode and a pressure node).

In one numerical approach to this problem via an FDM, we introduce a set of equally spaced nodes in the spatial dimension but leave the temporal dimension continuous. Suppose we divide the interval $0 \leq x \leq L$ into $N$ elements by $N+1$ nodes, each separated from its nearest neighbor(s) by the distance $\Delta x = L/N$. We denote the coordinates of those nodes by $x_i = i \Delta x$ with $i = 0, 1, 2, \ldots, N$. In particular, $x_0 = 0$ and $x_N = L$. Then, we introduce the set of functions $u_i(t) = u(x_i, t)$, evaluate the PDE of Eq. (15.132) at $x_i$, and express the spatial derivative at $x_i$ in terms of finite differences as in Eq. (15.75) to find that

$$\frac{d^2 u_i}{dt^2} = \frac{c^2}{\Delta x^2} \left( u_{i+1} - 2u_i + u_{i-1} \right)$$  \hspace{1cm} (15.135)
The boundary conditions can now be invoked to resolve problems with this equation when \( i = 0 \) and \( i = N \), at which points the equation refers to \( u \) at points outside the domain of the problem, i.e., to \( u_{-1} \) or \( u_{N+1} \), respectively. At \( x = 0 \), the boundary condition implies that \( u_0 \) = 0, so we can replace Eq. (15.135) at \( i = 0 \) with the equation \( u_0 = 0 \), though—for the sake of a similar treatment of all nodes—it is more appropriate to regard \( u_0 \) = 0 as the solution to the initial value problem

\[
\frac{d^2 u_0}{dt^2} = 0 \quad ; \quad u_0(0) = 0 \quad , \quad \frac{du_0}{dt}(0) = 0
\]  

(15.136)

To address Eq. (15.135) at \( i = N \), we look to a central difference approximation of the spatial derivative, finding that

\[
\frac{\partial u}{\partial x}(L, t) \approx \frac{u_{N+1}(t) - u_{N-1}(t)}{2\Delta x} = 0 \quad \Rightarrow \quad u_{N+1} = u_{N-1}
\]  

(15.137)

Then, we write Eq. (15.135) for \( i = N \), substitute \( u_{N-1} \) for \( u_{N+1} \), and solve the result for \( d^2 u_N/dt^2 \). In this approach, we have discretized only the spatial variable, and the task of solving a PDE in two variables becomes one of solving the coupled set

\[
\frac{d^2 u_0}{dt^2} = 0
\]  

(15.138)

\[
\frac{d^2 u_i}{dt^2} = \frac{c^2}{\Delta x^2} \left( u_{i+1} - 2u_i + u_{i-1} \right) \quad , \quad i = 1, 2, 3, \ldots, N - 1
\]  

(15.139)

\[
\frac{d^2 u_N}{dt^2} = \frac{c^2}{\Delta x^2} \left( -2u_N + 2u_{N-1} \right)
\]  

(15.140)

of ODEs subject to the initial conditions

\[
u_i(0) = f(x_i) \quad ; \quad \frac{du_i}{dt}(0) = g(x_i)
\]  

(15.141)

where, to reflect properly the conditions imposed on \( u_0 \), we suppose that \( f(x_0) = f(0) = 0 \) and \( g(x_0) = g(0) = 0 \). By discretizing the spatial variable, we have reduced our problem to a problem of the sort addressed in Chapter 11, and we therefore say no more about it here.

We could go one step further, discretizing also the time variable by introducing a time step \( \Delta t \) and the discrete time instants \( t_j = j \Delta t, \quad j = 0, 1, 2, \ldots. \) Then, approximating both the spatial and the temporal derivatives by finite differences and introducing the quantities \( u_{i,j} = u(x_i, t_j) \), we might evaluate the original PDE of Eq. (15.132) at \( (x, t) = (x_i, t_j) \) and find the fully discretized form

\[
\frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{\Delta t^2} = \frac{c^2}{\Delta x^2} \left( u_{i+1,j} - 2u_{i,j} + u_{i-1,j} \right)
\]  

(15.142)

Solving for \( u_{i,j+1} \) and introducing \( \alpha = c^2 \Delta t^2/\Delta x^2 \), we find the simpler form

\[
u_{i,j+1} = \alpha u_{i+1,j} + 2(1 - \alpha)u_{i,j} + \alpha u_{i-1,j} - u_{i,j-1}
\]  

(15.143)

Evidently, if we know values of \( u_{i,j} \) for \( j = 0 \) and \( j = 1 \), we can determine values for \( j = 2 \), then for \( j = 3, \ldots. \) That is, if we know \( u_{i,j} \) along two consecutive horizontal lines in Fig. 15.11, we can step the solution forward in time from line to line by increments of \( \Delta t \), going as far as our interest dictates and our patience endures.

As in the partial discretization of the previous paragraph, the boundary conditions help us deal with unknown values that appear when Eq. (15.143) is evaluated at \( i = 0 \) or \( i = N \) by stipulating that

\[
u_{0,j} = 0 \quad \text{and} \quad \frac{\partial u}{\partial x}(L, t_j) \approx \frac{u_{N+1,j} - u_{N-1,j}}{2\Delta x} = 0 \quad \Rightarrow \quad u_{N+1,j} = u_{N-1,j}
\]  

(15.144)
Figure 15.11: Geometry for full discretization of the wave equation. Equation 15.143 involves the points marked with solid figures and has been solved to express the solution at the (one) point marked with a solid square in terms of the solutions at the (four) points marked with solid circles.

Free of values outside the interval \( 0 \leq x \leq L \), the equations we use to step the solution to \( t_{j+1} \) from knowledge of the solution at previous times then are

\[
\begin{align*}
    u_{0,j+1} &= 0 \\
    u_{i,j+1} &= \alpha u_{i+1,j} + 2(1-\alpha)u_{i,j} + \alpha u_{i-1,j} - u_{i,j-1}, \quad i = 1, 2, \ldots, N-1 \\
    u_{N,j+1} &= 2\alpha u_{N-1,j} + 2(1-\alpha)u_{N,j} - u_{N,j-1}
\end{align*}
\]

In their turn, the initial conditions start the process by stipulating that

\[
\begin{align*}
    u_{i,0} &= f(x_i) \quad \text{and} \quad u_{i,1} = u_{i,0} + g(x_i) \Delta t
\end{align*}
\]

For consistency with the imposed boundary conditions, we must, of course, require that \( f(0) = 0 \) and \( dg(x)/dx|_{x=L} = 0 \). We leave it to Exercise 15.18 to demonstrate that the method described in this paragraph will be unstable unless \( \alpha \leq 1 \).

The discussion at the end of Section 15.2 on discretization and roundoff errors in finite difference methods applies as much in two dimensions as in one. Since solution of the wave equation involves a direct method, however, convergence error as discussed in that earlier section is not here an issue.

We can proceed no further in this example without resorting to a specific coding language, so we postpone further discussion to later sections.

### 15.16.2 The Diffusion Equation

As laid out in Section 15.1.2, the diffusion equation for heat flow in one dimension involves two independent variables, a spatial coordinate \( x \) locating a point in the one-dimensional medium and a time coordinate \( t \). The temperature \( u(x,t) \) at position \( x \) and time \( t \) in the medium satisfies Eq. (15.26),

\[
\frac{\partial u}{\partial t} = \alpha^2 \frac{\partial^2 u}{\partial x^2}
\]
which is to be solved subject to the initial condition

\[ u(x, 0) = f(x) \]  

that specifies the initial temperature distribution in the medium and to appropriate boundary conditions that specify the way the temperature is controlled at the two ends, say at \( x = 0 \) and \( x = L \).

To illustrate both possibilities in a single example, we suppose that, at \( x = 0 \), the temperature is maintained at the value \( T_0 \) and that, at \( x = L \), the medium is insulated so that no heat flow takes place either into or out of the rod at that end, i.e., we suppose that

\[ u(0, t) = T_0 \quad \text{and} \quad \frac{\partial u}{\partial x}(L, t) = 0 \]  

In one numerical approach to this problem via an FDM, we introduce a set of equally spaced nodes in the spatial dimension but leave the temporal dimension continuous. Suppose we divide the interval \( 0 \leq x \leq L \) into \( N \) elements by \( N + 1 \) nodes, each separated from its nearest neighbor(s) by the distance \( \Delta x = L/N \). We denote the coordinates of those nodes by \( x_i = i \Delta x \) with \( i = 0, 1, 2, \ldots, N \).

In particular, \( x_0 = 0 \) and \( x_N = L \). Then, we introduce the set of functions \( u_i(t) = u(x_i, t) \), evaluate the PDE of Eq. (15.149) at \( x_i \), and express the spatial derivative at \( x_i \) in terms of finite differences as in Eq. (15.75) to find that

\[ \frac{du_i}{dt} = \alpha^2 \frac{c}{\Delta x^2} \left( u_{i+1} - 2u_i + u_{i-1} \right) \]  

The boundary conditions can now be invoked to resolve problems with this equation when \( i = 0 \) and \( i = N \), at which points the equation refers to \( u \) at points outside the domain of the problem, i.e., to \( u_{-1} \) or \( u_{N+1} \), respectively. At \( x = 0 \), the boundary condition implies that \( u_0 = T_0 \), so we can replace Eq. (15.152) at \( i = 0 \) with the equation \( u_0 = T_0 \), though—for the sake of a similar treatment of all nodes—it is more appropriate to regard \( u_0 = T_0 \) as the solution to the initial value problem

\[ \frac{du_0}{dt} = 0 ; \quad u_0(0) = T_0 \]  

To address Eq. (15.152) at \( i = N \), we look to a central difference approximation of the spatial derivative, finding that

\[ \frac{\partial u}{\partial x}(L, t) \approx \frac{u_{N+1}(t) - u_{N-1}(t)}{2\Delta x} = 0 \quad \Rightarrow \quad u_{N+1} = u_{N-1} \]  

Then, we write Eq. (15.152) for \( i = N \), substitute \( u_{N-1} \) for \( u_{N+1} \), and solve the result for \( du_N/dt \). In this approach, in which we have discretized only the spatial variable, the problem of solving a PDE in two variables becomes one of solving the coupled set

\[ \frac{du_0}{dt} = 0 \]  

\[ \frac{du_i}{dt} = \frac{\alpha^2 c}{\Delta x^2} \left( u_{i+1} - 2u_i + u_{i-1} \right), \quad i = 1, 2, 3, \ldots, N - 1 \]  

\[ \frac{du_N}{dt} = 2\frac{\alpha^2 c}{\Delta x^2} \left( u_{N-1} - u_N \right) \]  

of ODEs subject to the initial condition

\[ u_i(0) = f(x_i) \]  

where, to reflect properly the conditions imposed on \( u_0 \), we suppose that \( f(x_0) = f(0) = T_0 \) and \( df(x)/dx|_{x=L} = 0 \). By discretizing the spatial variable, we have reduced our problem to a problem of the sort addressed in Chapter 11, and we therefore say no more about it here.
15.16. FINITE DIFFERENCE METHODS (FDMS) IN TWO DIMENSIONS

Figure 15.12: Geometry for full discretization of the diffusion equation. Equation 15.160 involves the points marked with solid figures and has been solved to express the solution at the (one) point marked with a solid square in terms of the solutions at the (three) points marked with solid circles.

We could go one step further, discretizing also the time variable by introducing a time step $\Delta t$ and the discrete time instants $t_j = j \Delta t$, $j = 0, 1, 2, \ldots$. Then, approximating both the spatial and the temporal derivatives by finite differences and introducing the quantities $u_{ij} = u(x_i, t_j)$, we might evaluate the original PDE of Eq. (15.149) at $(x, t) = (x_i, t_j)$ and find the fully discretized form

$$
\frac{u_{i,j+1} - u_{i,j}}{\Delta t} = \frac{\alpha^2}{\Delta x^2} \left( u_{i+1,j} - 2u_{i,j} + u_{i-1,j} \right) \quad (15.159)
$$

Solving for $u_{i,j+1}$ and introducing $\gamma = \frac{\alpha^2 \Delta t}{\Delta x^2}$, we find the simpler form

$$
u_{i,j+1} = \gamma u_{i-1,j} + (1 - 2\gamma)u_{i,j} + \gamma u_{i+1,j} \quad (15.160)
$$

Evidently, if we know values of $u_{i,j}$ for $j = 0$, we can determine values for $j = 1$, then for $j = 2, \ldots$. That is, if we know $u_{i,j}$ along one horizontal line in Fig. 15.12, we can step the solution forward in time from line to line by increments of $\Delta t$, going as far as our interest dictates and our patience endures.

As in the partial discretization of the previous paragraph, the boundary conditions help us deal with unknown values that appear when Eq. (15.160) is evaluated at $i = 0$ or $i = N$ by stipulating that

$$u_{0,j} = T_0 \quad \text{and} \quad \frac{\partial u}{\partial x}(L,t_j) \approx \frac{u_{N+1,j} - u_{N-1,j}}{2 \Delta x} = 0 \quad \implies \quad u_{N+1,j} = u_{N-1,j} \quad (15.161)
$$

Free of values outside the interval $0 \leq x \leq L$, the equations we use to step the solution to $t_{j+1}$ from knowledge of the solution at previous times then are

$$
u_{0,j+1} = T_0 \quad (15.162)
$$

$$
u_{i,j+1} = \gamma u_{i-1,j} + (1 - 2\gamma)u_{i,j} + \gamma u_{i+1,j}, \quad i = 1, 2, \ldots, N - 1 \quad (15.163)
$$

$$u_{N,j+1} = 2\gamma u_{N-1,j} + (1 - 2\gamma)u_{N,j} \quad (15.164)
$$

In their turn, the initial conditions start the process by stipulating that

$$u_{i,0} = f(x_i) \quad (15.165)$$
For consistency with the imposed boundary conditions, we must, of course, require that $f(0) = T_0$. We leave it to Exercise 15.19 to demonstrate that the method described in this paragraph will be unstable unless $\gamma \leq 1/2$.

The discussion at the end of Section 15.2 on discretization error and roundoff errors in finite difference methods applies as much in two dimensions as in one. Since solution of the diffusion equation involves a direct method, however, convergence error as discussed in that earlier section is not here an issue.

We can proceed no further in this example without resorting to a specific coding language, so we postpone further discussion to later sections.

### 15.16.3 The Laplace Equation

As laid out in Section 15.1.3, the Laplace equation for the steady state temperature distribution in a two-dimensional plate involves two independent spatial variables $x$ and $y$ which, together, locate a point in the two-dimensional medium. The temperature $u(x,y)$ at the point $(x,y)$ satisfies Eq. (15.27),

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \tag{15.166}$$

which is to be solved subject to boundary conditions that specify the way the temperature is controlled at the boundaries of the region.\(^3^0\)\(^3^1\)

Stipulation of boundary conditions is more complicated for a problem involving two spatial dimensions than for the problems in Sections 15.16.1 and 15.16.2, each of which involved only one spatial dimension. In one spatial dimension, the boundary consists of two points, one at each end of the region. In two spatial dimensions, the boundary is defined by a closed curve—whatever is necessary to bound the region of interest—in the $xy$ plane. Easy application of FDMs is limited to fairly simple geometries in which the boundaries lie along coordinate lines in one or another of the standard coordinate systems. In the present example, we will suppose the region of interest to be a square, each of whose edges has length $L$, and we will suppose that its four corners lie at $(x,y) = (0,0), (L,0), (L,L), \text{ and } (0,L)$ in the $xy$ plane. Then, so as to illustrate both possible types of boundary condition, we will suppose that

$$u(x,0) = 0 \quad ; \quad u(x,L) = 100 \quad ; \quad u(0,y) = 100 \frac{y}{L} \quad ; \quad \frac{\partial u}{\partial x}(L,y) = 0 \tag{15.167}$$

i.e., that the temperature along the line $y = 0$ is maintained at the value 0, that the temperature along the line $y = L$ is maintained at the value 100, that the temperature along the line $x = 0$ rises linearly from 0 at $y = 0$ to 100 at $y = L$, and that the edge along the line $x = L$ is insulated.

The partial discretization of a PDE as illustrated in the previous two sections is applicable only to initial value problems and does not provide a suitable approach to the present boundary value problem. We can approach the present problem successfully only through full discretization of the PDE. Thus, we divide each edge into $N$ segments each of length $\Delta x = L/N$, placing the $(N+1)^2$ nodes at the points $(x_i,y_j)$, where $x_i = i \Delta x$ ($i = 0,1,2,\ldots,N$) and $y_j = j \Delta y$ ($j = 0,1,2,\ldots,N$). Then, we introduce the $(N+1)^2$ values $u_{i,j} = u(x_i,y_j)$, evaluate Eq. (15.166) at the point $(x_i,y_j)$, and express each second partial derivative in terms of finite differences to find that

$$\frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{\Delta x^2} + \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{\Delta x^2} = 0 \tag{15.168}$$

\(^3^0\)In this case, there are no initial conditions, since we seek only the final steady state temperature, which is determined solely by the (time-independent) boundary conditions. The initial temperature distribution is entirely irrelevant.

\(^3^1\)Note that this approach to the Laplace equation was also discussed in Section 9.3.1, though the example discussed there involved only Dirichlet boundary conditions.
15.16. FINITE DIFFERENCE METHODS (FDMS) IN TWO DIMENSIONS

Figure 15.13: Geometry for full discretization of the Laplace equation. Equation 15.169 involves the points marked with solid figures and has been solved to express the solution at the (one) point marked with a solid square in terms of the solutions at the (four) points marked with solid circles.

or, solving for $u_{i,j}$, that

$$u_{i,j} = \frac{1}{4} \left( u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} \right)$$  \hspace{1cm} (15.169)

Interestingly, in solutions to Laplace’s equation via an FDM, the value at a particular node is equal to the average of the values at the four nearest neighbors of that node. This geometry is shown in Fig. 15.13.

The boundary conditions can now be invoked to resolve problems with Eq. (15.169) when $i = 0$ or $i = N$ and/or $j = 0$ or $j = N$, at which points the equation refers to one or more nodes lying outside the boundaries of the region in which a solution is sought. We must deal with eight different categories of such nodes:

- nodes on the bottom edge but not at a corner, for which $j = 0$ and $i = 1, 2, \ldots, N - 1$. Along this edge, the boundary condition stipulates that $u_{i,0} = 0$, and we replace Eq. (15.169) with this alternative.

- nodes on the top edge but not at a corner, for which $j = N$ and $i = 1, 2, \ldots, N - 1$. Along this edge, the boundary condition stipulates that $u_{i,0} = 100$, and we replace Eq. (15.169) with this alternative.

- nodes on the left edge but not at a corner, for which $i = 0$ and $j = 1, 2, \ldots, N - 1$. Along this edge, the boundary condition stipulates that $u(0,y) = 100y/L$, or $u_{0,j} = 100y_j/L$, and we replace Eq. (15.169) with this alternative.

- nodes on the right edge but not at a corner, for which $i = N$ and $j = 1, 2, \ldots, N - 1$. This edge is to be insulated, i.e., $\partial u/\partial x = 0$ on that edge. We use a central difference approximation to the derivative to find that

$$\frac{\partial u}{\partial x}(x,y) \bigg|_{x=L} = 0 \implies \frac{u_{N+1,j} - u_{N-1,j}}{2 \Delta x} = 0 \implies u_{N+1,j} = u_{N-1,j}$$  \hspace{1cm} (15.170)
For these nodes, Eq. (15.169) becomes

\[ u_{N,j} = \frac{1}{4} \left( u_{N+1,j} + u_{N-1,j} + u_{N,j+1} + u_{N,j-1} \right) \]

\[ = \frac{1}{4} \left( 2u_{N-1,j} + u_{N,j+1} + u_{N,j-1} \right) \]  

(15.171)

and, as long as \( j \) stays in the specified range, this equation no longer involves points outside the region defined by the boundaries.

- the node at the lower left corner for which \( i = 0 \) and \( j = 0 \). This point lies on two of the bounding edges. Fortunately, in the present case, the boundary conditions on those two edges are consistent, and we simply replace Eq. (15.169) with \( u_{0,0} = 0 \). In fact, however, this point will never enter into any equation we will need to consider, so the value we assign at this one point is of no consequence. Indeed, in some problems, there may be a discontinuity in the temperature at an isolated point such as this one. Conveniently, the method we have adopted is not upset by such a discontinuity.

- the node at the upper left corner, for which \( i = 0 \) and \( j = N \). This point is similar to the point \((i,j) = (0,0)\), even to the consistency of the values on the two edges to which it belongs. We set \( u_{0,N} = 100 \).

- the node at the lower right corner, for which \( i = N \) and \( j = 0 \). Here, the boundary condition on the lower edge dictates that we should set \( u_{N,0} = 0 \). Applied to this node, however, Eq. (15.171) with \( j = 0 \) suggests that we should require that

\[ u_{N,0} = \frac{1}{4} \left( 2u_{N-1,0} + u_{N,1} + u_{N,-1} \right) \]

(15.172)

Consistency with the presumed value \( u_{N,0} = 0 \) and the known value \( u_{N-1,0} = 0 \) then implies that we should expect that

\[ u_{N,-1} = -u_{N,1} \]  

(15.173)

a result that we might also infer from symmetries if we saw the problem of interest as the upper half of a larger problem obtained by reflecting the problem we are addressing into the region \(-L < y < 0\), making the temperature on the bottom edge of that larger region \(-100\). We conclude that taking \( u_{N,0} = 0 \) is appropriate and justified.

- the node at the upper right corner, for which \( i = N \) and \( j = N \). This point is similar to the point \((i,j) = (N,0)\) and, without further argument, we accept the replacement \( u_{N,N} = 100 \).

In short, the equations we seek to solve, now involving no quantities at points outside the boundaries of the problem, are

\[ u_{i,j} = \frac{1}{4} \left( u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} \right) \quad ; \quad 1 \leq i, j \leq N - 1 \]

\[ u_{i,0} = 0 \quad ; \quad 1 \leq i \leq N - 1 \]

\[ u_{i,N} = 100 \quad ; \quad 1 \leq i \leq N - 1 \]

\[ u_{0,j} = 100 \frac{y_j}{L} \quad ; \quad 1 \leq j \leq N - 1 \]

\[ u_{N,j} = \frac{1}{4} \left( 2u_{N-1,j} + u_{N,j+1} + u_{N,j-1} \right) \quad ; \quad 1 \leq j \leq N - 1 \]

\[ u_{0,0} = 0 \]

\[ u_{0,N} = 100 \]

\[ u_{N,0} = 0 \]

\[ u_{N,N} = 100 \]

(15.174)
We can proceed no further in this example without resorting to a specific coding language, so we postpone further discussion to later sections. We do note here, however, that iterative methods will almost always be used to solve this set of equations. Consequently (and in contrast to the situation with the wave and diffusion equations), we will have to pay attention not only to discretization and roundoff errors but also to convergence error.

15.19 Using OCTAVE to Solve 2D PDEs via an FDM

To complete the solution of the problems laid out in Section 15.16, we must choose a specific language for the development of explicit coding to implement the algorithms described. In this section, we illustrate that process in OCTAVE.

15.19.1 The Wave Equation

The final step in addressing the example laid out in Section 15.16.1 is to set up and solve Eqs. (15.145)–(15.147). Using a command file in OCTAVE, we would begin by requesting input of all necessary parameters and assuring that each is stored with the proper data type. The statements

\[
\begin{align*}
N &= \text{input( 'Enter number of segments (N): ' ); } N = \text{fix}(N); \\
\text{dt} &= \text{input( 'Enter time step (dt): ' );} \\
T &= \text{input( 'Enter number of time steps (T): ' ); } T = \text{fix}(T); \\
c &= \text{input( 'Enter speed of propagation (c): ' );} \\
L &= \text{input( 'Enter length of string (L): ' );}
\end{align*}
\]

accomplish those objectives. Then, we would determine the length of each segment, establish values for the coordinates at which solutions will be generated, and evaluate and display the one parameter \( \alpha \) with the statements

\[
\begin{align*}
dx &= L/N; \\
x &= dx*[0:N]; \\
alpha &= c^2*dt^2/dx^2; \\
\text{alpha}
\end{align*}
\]

Prudence dictates the wisdom of adding the statement

\[
\begin{align*}
\text{if alpha} &> 1.0 \\
\quad \text{disp( 'Error: alpha} > 1; \text{ execution halted' );} \\
\quad \text{return;}
\end{align*}
\]

if alpha > 1.0
    disp( 'Error: alpha > 1; execution halted' );
    return;
end

To halt execution if the parameter \( \alpha \) has a value that guarantees an unstable—and hence inaccurate—solution.

As we saw in the general discussion in Section 15.16.1, generation of the solution at the next time instant requires knowledge of the solution at the current time instant and of the solution at the immediate past time instant. At each step of the way, we need preserve only the current and immediate past solutions, but we must keep those solutions until the solution at the next time instant has been generated. In essence, we need three vectors of dimension \( N+1 \) for storing solutions. We declare, therefore, that at any step in the process, the vector \( \textbf{u1} \) will store the past solution, the vector \( \textbf{u2} \) will store the current solution, and the vector \( \textbf{u3} \) will receive the solution at the next time
instant as it is generated. Then, once the new solution has been generated and displayed (graphed or written to a file), we no longer need the values in $u_1$, so we will move those in $u_2$ to $u_1$ and those in $u_3$ to $u_2$, thus preparing for the next pass through a loop that advances the solution from time instant to time instant. These three variables are prepared with the statements

\[
\begin{align*}
\text{u1} &= \text{zeros}(N+1,1); & \text{\% For past solution} \\
\text{u2} &= \text{zeros}(N+1,1); & \text{\% For current solution} \\
\text{u3} &= \text{zeros}(N+1,1); & \text{\% For next solution}
\end{align*}
\]

Next, before coding the loop that will generate the solution, step by step, we must initialize the values in $u_1$ to reflect the initial displacement of the string, initialize the values in $u_2$ to reflect the impact of the initial velocity on the motion during the first time step, and display $u_1$ and $u_2$. To be specific, let us suppose that, initially, the string is displaced in the shape of a single hump of a sine wave, but only in the middle quarter of its length, and is not displaced over the first three-eighths and the last three eighths of its length. Thus, we suppose that

\[
f(x) = \begin{cases} 
0 & 0 \leq x \leq \frac{3}{8}L \\
1 + \cos \frac{8\pi}{L} \left( x - \frac{L}{2} \right) & \frac{3}{8}L \leq x \leq \frac{5}{8}L \\
0 & \frac{5}{8}L \leq x \leq L 
\end{cases}
\]  

(15.175)

which results in a smooth transition from zero displacement outside the center one quarter of the string and the sinusoidal displacement in that interval. Further, we suppose that the string is released from rest so that

\[
g(x) = 0
\]  

(15.176)

or, equivalently—see Eq. (15.148)—$u_{i,0} = u_{i,1}$ or $u_2 = u_1$. The coding that will impose these initial conditions, display the initial solution, and then display the solution after the first time step then is\textsuperscript{32}

\[
\begin{align*}
b &= 8.0 \pi / L \\
\text{for } i &= \text{fix}(3\times N/8) + 1 : \text{fix}(5\times N/8) \\
& \quad \text{u1}(i) = 1.0 + \cos(b \times (x(i)-L/2.0)); \\
\end{align*}
\]

plot( x, u1, 'linewidth', 3, 'color', 'black' )
set(gca, 'ylim', [-2.0,2.0] );

Here, we have recognized that the displacement over time will range from $-2.0$ to $+2.0$, and we have set the vertical scale to display that range. We will need to include that specification for each subsequent invocation of the command plot. Next, because zero initial velocity means that the solution at the second time is the same as that at the first time, we execute the statements

\[
\begin{align*}
\text{u2} &= \text{u1}; \\
\text{plot( x, u2, 'linewidth', 3.0, 'color', 'black' ) } \\
\text{set(gca, 'ylim', [-2.0,2.0] );}
\end{align*}
\]

to take the first step and plot the solution at that step.

\textsuperscript{32}The calculation of the range of i to be used is complicated. The OCTAVE function \texttt{fix} will truncate the quantities 3\times N/8 and 5\times N/8. We really want the lower limit to be raised rather than truncated. The upper limit can be truncated. Thus, we add 1 to the lower limit.
for j=2:T
    u3(1) = 0;
    for i=2:N
        u3(i) = alpha*u2(i+1) + ...
        2.0*(1.0-alpha)*u2(i) + alpha*u2(i-1) - u1(i);
    end
    u3(N+1) = 2*alpha*u2(N)+2*(1.0-alpha)*u2(N+1) - u1(N+1);
    u1 = u2; u2 = u3;
    plot( x, u2, 'linewidth', 3, 'color', 'black' )
    set(gca, 'ylim', [-2.0,2.0] );
end

Here, each pass through the outermost loop advances the solution by one time step. Within that loop, we (1) construct the solution at the next time instant by exploiting Eqs. (15.145)–(15.147), (2) plot the solution, and (3) shift the values to prepare for the next pass through the loop. A more fully commented command file containing these statements—and a few others to be discussed in a moment—is named fdmwave1d.m, is listed in Appendix 15.D.1, and can be copied from the directory $HEAD/octave.33

Two problems emerge when we run the coding developed to this point with a trial set of values, say $N = 100$, $dt = 0.1$, $T = 20$, $c = 0.5$, and $L = 10$ (for which $\alpha = 0.25$). First, the solution is generated and displayed so quickly that we cannot follow its evolution. Second, even with 20 time steps, the solution is not advanced very far.

We fix the the first of these problems by introducing a time delay after each plot is produced by inserting the statement

```
pause( delay )
```

after each `plot` statement and inserting the statements

```matlab
delay = input( 'Delay between plots in secs (delay): ' );
```

at the beginning so the user can specify the desired delay at execution time.34

We can, of course, fix the second problem simply by requesting a larger number of time steps or enlarging the time step (or both—though we must be careful not to increase $\alpha$ beyond 1.0). As the number of time steps requested increases—and, perhaps, the time interval between steps decreases—the program as we now have it, which displays the solution at every step of the way, will produce increasing volumes of output. With the more accurate solutions (numerous segments, short time interval between generated solutions, many time steps to be computed), we would be wise to introduce a mechanism for suppressing the display of many of the generated solutions. To do so, we add at the beginning the statement

```
f = input( 'Plot frequency (f): ' ); f = fix(f);
```

to request the number of solutions whose display should be suppressed after a particular solution has been displayed and, in addition, we modify the `plot` statement in the final loops to make the display conditional upon $j$—the loop index—being a multiple of $f$. In effect, we simply condition the all `plot` statements except the first on in a way that allows the `each` step to be calculated but only every $f$-th step to be displayed. The second `plot` statement, including the associated `set` and `pause` statements will be executed only if $f = 1$, i.e., `all` steps are to be displayed. All of the remaining `plot`, `set`, and `pause` statements will be recast in the form35

---

33At some sites, the file may also be located in the OCTAVE directory structure such that OCTAVE can find it when it is identified only by its name without a prepended path.
34In OCTAVE, the `pause` command is by default enabled.
35We exploit here the fact that the quantity $f*\text{fix}(j/f)$ will be equal to $j$ only when $j$ is a multiple of $f$. 

if f*fix(j/f) == j
    plot( x, u2, 'linewidth', 3, 'color', 'black' );
    pause( delay );
end

(See the full listing in Appendix 15.D.1 for the details.)

With these embellishments, we might run the program and—with a bit of trial and error—submit the controlling values

fdmwave1d
Enter number of segments (N): 100
Enter time step (dt): 0.1
Enter number of time steps (T): 800
Enter speed of propagation (c): 0.5
Enter length of string (L): 10
Delay between plots in secs (delay): 0.1
Plot frequency (f): 2
alpha = 0.2500

By examining the output, we conclude that these parameters advance the solution through what we might anticipate to be one full period of the motion. The process evidently then repeats, since this first cycle has returned the string (essentially—recognize we are generating an approximate solution) to its initial configuration.

Several representative configurations of the string are shown in Fig. 15.14. The pulse divides into two equal halves, one propagating to the left and the other to the right. Each reflects at the end, though with a phase change at the fixed (left) end and without a phase change at the free (right) end. The pulses pass through one another, create at $t = 20.0$ s an instant when—except for discretization error and/or computer roundoff error—the string is straight but its center portion has non-zero velocity, reflect from the ends again, and return to the middle, this time in phase. At $t = 40.0$ s, half the cycle has been completed and the original shape has been inverted. In another 40.0 s (i.e. at $t = 80.0$ s), the string will return to its initial configuration. Note in particular the amplitude of the motion of the free end of the string when the pulse reflects there. Because the reflected pulse is in phase with the incident pulse, the two add constructively and, when half the pulse has reflected, the displacement at the end is twice that of the incident pulse. At the fixed end, the reflected pulse is out of phase with the incident pulse, so their interference at all instants during the reflection is completely destructive at $x = 0$, and the end remains fixed.

Accuracy is difficult to assess in this situation. Given the speed of propagation to be 0.5 m/s and the length of the string to be 10 m, we know that each (half) pulse should take 10 s (5 m/0.5 m/s) to travel from its initial position to the end of the string, 20 s to return to the middle of the string, etc. We also know that the two pulses, now of opposite polarity, should superpose at $t = 20$ s and exactly cancel each other, should superpose at $t = 40$ s to exactly invert the initial pulse, . . . . The solutions shown in Fig. 15.14 at these times reflect these expectations reasonably well. To check one of these expectations out more fully, let us run fdmwave1d to obtain values for the solution at $t = 20$ s for three different accuracies. To recreate the frame at $t = 20$ s in Fig. 15.14, we would execute the statements and provide the input

fdmwave1d
Enter number of segments (N): 100
Enter time step (dt): 0.1
Enter number of time steps (T): 200
Enter speed of propagation (c): 0.5
Enter length of string (L): 10
Figure 15.14: Representative frames in motion of a string. This graph was produced with OCTAVE.

Delay between plots in secs (delay): 0
Plot frequency (f): 20
alpha = 0.2500
u2_100 = u2; x_100 = x;
[ max(u2_100), min(u2_100) ]
ans = 0.12434, -0.12415

Here, at the next to the last OCTAVE prompt, we save the final values for later plotting and, in the last OCTAVE prompt, we display the range of values assumed by the solution at t = 20 s. Then, with the input,

```
fdmwave1d
Enter number of segments (N): 200
Enter time step (dt): 0.05
Enter number of time steps (T): 400
Enter speed of propagation (c): 0.5
Enter length of string (L): 10
Delay between plots in secs (delay): 0
Plot frequency (f): 40
alpha = 0.2500
u2_200 = u2; x_200 = x;
[ max(u2_200), min(u2_200) ]
ans = 0.043713, -0.043697
```
and again with the input

```
fdmwaveld
Enter number of segments (N): 400
Enter time step (dt): 0.025
Enter number of time steps (T): 800
Enter speed of propagation (c): 0.5
Enter length of string (L): 10
Delay between plots in secs (delay): 0
Plot frequency (f): 200
alpha = 0.2500
u2_400 = u2; x_400 = x;
[ max(u2_400), min(u2_400) ]
ans = 0.019038 -0.019036
```

we generate and save the solution at \( t = 20 \) s, using progressively smaller divisions of the string (larger \( N \)), adjusting the other parameters to track the solution precisely to \( t = 20 \) s, omitting numerous intermediate graphs, preserving the parameter \( \alpha \) at the value 0.25, and displaying the maximum and minimum values attained by the approximate solution (which we expect to be zero everywhere). Significantly, the departure from zero changes from \( \pm 0.12 \) to \( \pm 0.04 \) to \( \pm 0.02 \) as the number of segments increases—and presumably the accuracy of the solution increases. For graphical comparison (and choosing a vertical scale to reveal the differences most clearly), we generate Fig. 15.15 with the statements

```
subplot(3,1,1)
plot( x_100, u2_100, 'linewidth', 3, 'color', 'black' )
title( 'N = 100', 'fontsize', 14 )
set(gca, 'YLim', [-0.2,0.2] )
subplot(3,1,2)
plot( x_200, u2_200, 'linewidth', 3, 'color', 'black' )
title( 'N = 200', 'fontsize', 14 )
set(gca, 'YLim', [-0.2,0.2] )
subplot(3,1,3)
plot( x_400, u2_400, 'linewidth', 3, 'color', 'black' )
title( 'N = 400', 'fontsize', 14 )
set(gca, 'YLim', [-0.2,0.2] )
```

Increasing \( N \) clearly improves the solution, i.e., makes it more nearly zero everywhere. Given these results, we might find it wise to go back and regenerate Fig. 15.14 with \( N = 400 \).

### 15.19.2 The Diffusion Equation

The final step in addressing the example laid out in Section 15.16.2 is to set up and solve Eqs. (15.162)–(15.164). Using a command file in OCTAVE, we would begin by requesting input of all necessary parameters and assuring that each is stored with the proper data type. The statements

```
N = input('Enter number of segments (N): '); N = fix(N);
dt = input('Enter time step (dt): ');
T = input('Enter number of time steps (T): '); T = fix(T);
alpha = input('Enter value of alpha (alpha): ');
L = input('Enter length of rod (L): ');
```
Figure 15.15: Shape of the string at $t = 20$ s for different segment sizes. This graph was produced with OCTAVE.

accomplish those objectives. Then, we would determine the length of each segment, establish values for the coordinates at which solutions will be generated, and evaluate and display the one parameter $\gamma$ with the statements

$$\begin{align*}
    dx &= L/N; \\
    x &= dx*[0:N]; \\
    gamma &= alpha^2*dt/dx^2; \\
    gamma
\end{align*}$$

Prudence dictates the wisdom of adding the statement

$$\begin{align*}
    if gamma > 0.5
        disp('Error: gamma > 0.5; execution halted'); \\
        return;
    end
\end{align*}$$

to halt execution if the parameter $\gamma$ has a value that guarantees an unstable—and hence inaccurate—solution.

As we saw in the general discussion in Section 15.16.2, generation of the solution at the next time instant requires knowledge only of the solution at the current time instant. At each step of the way, we need preserve only the current solution, but we must keep that solution until the solution at the next time instant has been generated. In essence, we need two vectors of dimension $N + 1$ for storing solutions. We declare, therefore, that at any step in the process, the vector $u1$ will store the current solution, and the vector $u2$ will receive the solution at the next time instant as it is generated. Then, once the new solution has been generated and displayed (graphed or written to a
file), we no longer need the values in $u_1$, so we will move the solution in $u_2$ to $u_1$, thus setting the stage for the next pass through a loop that advances the solution from time instant to time instant. These two variables are prepared with the statements

\begin{verbatim}
u1 = zeros( N+1, 1 ); % For current solution
u2 = zeros( N+1, 1 ); % For next solution
\end{verbatim}

Next, before coding the loop that will generate the solution, step by step, we must initialize the values in $u_1$ to reflect the initial temperature distribution in the rod and then display that distribution. To be specific, let us suppose that, initially, the temperature varies like one hump of a sine wave, but only in the middle quarter of its length, and is zero over the first three-eighths and the last three eighths of its length. Thus, we suppose that

\[ f(x) = \begin{cases} 
0 & 0 \leq x \leq \frac{3}{8}L \\
1 + \cos \frac{8\pi}{L} \left( x - \frac{L}{2} \right) & \frac{3}{8}L \leq x \leq \frac{5}{8}L \\
0 & \frac{5}{8}L \leq x \leq L 
\end{cases} \]  

(15.177)

which results in a smooth transition from zero temperature outside the center one quarter of the string to the sinusoidal displacement in that interval. The coding that will impose these initial conditions and then display the initial solution is

\begin{verbatim}
b = 8.0*pi/L;
for i=fix(3*N/8)+1:fix(5*N/8)
    u1(i) = 1.0 + cos(b*(x(i)-L/2.0));
end
plot( x, u1, 'linewidth', 3, 'color', 'black' )
title('t = 0 s', 'fontsize',14)
set(gca, 'YLim', [ 0.0, 2.0 ] )
\end{verbatim}

Here, we have recognized that the temperature will always remain between 0 and 2, and we have set the vertical scale to display that range for each anticipated graph.

Now, taking the temperature at $x = 0$ to be zero ($T_0 = 0$), we are ready to code the algorithm that uses Eqs. (15.162)–(15.164) to advance the solution, step by step. Appropriate coding, or at least a first pass at such coding, might be

\begin{verbatim}
for j=1:T
    u2(1) = 0.0;
    for i=2:N
        u2(i) = gamma*u1(i-1) + (1.0-2.0*gamma)*u1(i) + gamma*u1(i+1);
    end
    u2(N+1) = 2*gamma*u1(N) + (1-2*gamma)*u1(N+1);
    plot( x, u1, 'linewidth', 3, 'color', 'black' )
    title( [ 't = ', num2str(dt*j), ' s' ], 'fontsize',14)
    set(gca, 'YLim', [ 0.0, 2.0 ] )
    u1 = u2;
end
\end{verbatim}

The calculation of the range of $i$ to be used is complicated. Because of integer arithmetic, the quantities $3*8/8$ and $5*8/8$ will be truncated, but we really want the lower limit to be raised rather than truncated. The upper limit can be truncated. Thus, we add 1 to the lower limit.
Here, each pass through the outermost loop advances the solution by one time step. Within that loop, we (1) construct the solution at the next time instant by exploiting Eqs. (15.162)–(15.164), (2) plot the solution, and (3) shift the values to prepare for the next pass through the loop. A more fully commented command file containing these statements—and a few others to be discussed in a moment—is named \texttt{fdmdiffus1d.m}, is listed in Appendix 15.D.2, and can be copied from the directory \texttt{$\$\text{HEAD/octave}}.\footnote{At some sites, the file may also be located in the OCTAVE directory structure such that OCTAVE can find it when it is identified only by its name without a prepended path.}

Two problems emerge when we run the coding developed to this point with a trial set of values, say $N = 100$, $dt = 0.1$, $T = 25$, $\alpha = 0.1$, and $L = 10$ (for which $\gamma = 0.1$). First, the solution is generated and displayed so quickly that we cannot follow its evolution. Second, even with 25 time steps, the solution is not advanced very far.

We fix the first of these problems by introducing a time delay after each plot is produced by inserting the statement

\begin{verbatim}
    pause( delay );
\end{verbatim}

after each \texttt{plot} statement and inserting the statement

\begin{verbatim}
    delay = input( 'Delay between plots in secs (delay): ' );
\end{verbatim}

at the beginning so the user can specify the desired delay at execution time.

We can, of course, fix the second problem simply by requesting a larger number of time steps or by enlarging the time step (or both—though we must be careful not to increase $\gamma$ beyond 0.5). As the number of time steps requested increases, however, the program as we now have it will display the solution at every step of the way. With the more accurate solutions (numerous segments, short time interval between generated solutions, many time steps to be computed), we would be wise to introduce a mechanism for suppressing the display of many of the generated solutions. To do so, we add at the beginning the statement

\begin{verbatim}
    f = input( 'Plot frequency (f): ' ); f = fix(f);
\end{verbatim}

to request the number of solutions whose display should be suppressed after a particular plot has been produced and, in addition, we modify the \texttt{plot} statement in the final loops to make the display conditional upon $j$—the loop index—being a multiple of $f$. In effect, we simply replace the \texttt{plot} statement everywhere it appears with the conditional statement

\begin{verbatim}
    if f*fix(j/f) == j
        plot( x, u1, 'linewidth', 3, 'color', 'black' )
        set(gca, 'YLim', [ 0.0, 2.0 ])
        pause( delay );
    end
\end{verbatim}

(See the full listing in Appendix 15.D.2 for the details.)

With this embellishment, we might run the program and—with a bit of trial and error—submit the starting values

\begin{verbatim}
fdmdiffus1d
Enter number of segments (N): 100
Enter time step (dt): 0.4
Enter number of time steps (T): 4400
\end{verbatim}
Figure 15.16: Temperatures in a rod at various times as it approaches equilibrium. Note that the vertical scale after the graph for \( t = 0.0 \) s has been expanded the better to show the gradual evolution of the temperature distribution. This graph was produced with OCTAVE.

Enter value of alpha (\( \alpha \)): 0.1
Enter length of rod (L): 10
Delay between plots in secs (delay): 0.1
Plot frequency (f): 50
gamma = 0.4000

to trace the evolution of the initial temperature distribution for a longer time.

Several representative graphs of the temperature distribution versus position are shown in Fig. 15.16. Initially, the thermal energy in the central peak diffuses to either and of the rod. Energy diffuses out of the rod altogether at \( x = 0 \), because the “refrigerator” that maintains that end at \( u = 0 \) °C functions as a sink for thermal energy. Any thermal energy that reaches the insulated end \( (x = 10) \), however, does not escape the rod. In time, the insulated end becomes the hottest part of the rod. Then, thermal energy diffuses from the insulated end back towards the refrigerated end. Ultimately, all of the initial thermal energy in the rod has “leaked” into the refrigerator and the rod has attained equilibrium. At that point, the entire rod has the temperature of the refrigerator. Evidently, the insulated end is hottest at about \( t = 1200 \) s. Thereafter, the temperature at all points in the rod slowly falls until all points reach the temperature of the refrigerator. We might well have been able to predict this end result, even without solving the problem explicitly.

To assess the accuracy of this solution, let us generate and compare solutions at \( t = 160 \) s for \( N = 100, 200, \) and 400. To recreate the frame at \( t = 160 \) s in Fig. 15.16, we would execute the statements and provide the input.
15.19. USING OCTAVE TO SOLVE 2D PDES VIA AN FDM

fdmdiffus1d
Enter number of segments (N): 100
Enter time step (dt): 0.4
Enter number of time steps (T): 400
Enter value of alpha (alpha): 0.1
Enter length of rod (L): 10
Delay between plots in secs (delay): 0.1
Plot frequency (f): 25
gamma = 0.4000
u1_100 = u1; x_100 = x;

Here, in the last line, we save the final values for later plotting. Then, with the input,

fdmdiffus1d
Enter number of segments (N): 200
Enter time step (dt): 0.1
Enter number of time steps (T): 1600
Enter value of alpha (alpha): 0.1
Enter length of rod (L): 10
Delay between plots in secs (delay): 0.1
Plot frequency (f): 100
gamma = 0.4000
u1_200 = u1; x_200 = x;

and again with the input,

fdmdiffus1d
Enter number of segments (N): 400
Enter time step (dt): 0.025
Enter number of time steps (T): 6400
Enter value of alpha (alpha): 0.1
Enter length of rod (L): 10
Delay between plots in secs (delay): 0.1
Plot frequency (f): 400
gamma = 0.4000
u1_400 = u1; x_400 = x;

we generate and save the solution at \( t = 160 \) s, using progressively smaller divisions of the rod (larger \( N \)), adjusting the other parameters to track the solution precisely to \( t = 160 \) s, preserving the parameter \( \gamma \) at the value 0.4.\(^{38}\) For graphical comparison (and choosing a vertical scale to reveal the differences most clearly), we generate Fig. 15.17 with the statements

```octave
subplot(3,1,1)
plot( x_100, u1_100, 'linewidth', 3, 'color', 'black' )
title( 'N = 100', 'fontsize', 14 )
set(gca, 'YLim', [0.0,0.6] )
subplot(3,1,2)
plot( x_200, u1_200, 'linewidth', 3, 'color', 'black' )
title( 'N = 200', 'fontsize', 14 )
set(gca, 'YLim', [0.0,0.6] )
subplot(3,1,3)
```

\(^{38}\)Note that, since \( \gamma = \alpha^2 dt/dx^2 \), doubling \( N \) which halves \( dx \) and reduces \( dx^2 \) to one quarter of its original value requires that \( dt \) be reduced by a factor of four as well if \( \gamma \) is to retain its original value.
CHAPTER 15. SOLVING PARTIAL DIFFERENTIAL EQUATIONS

Figure 15.17: Temperature distribution at $t = 160$ s for different segment sizes. This graph was produced with OCTAVE.

The three frames in Fig. 15.17 appear to be identical. Indeed, were the solutions over plotted, no difference—to the resolution of the graph—would be noticed anywhere along the rod. Evidently, at $t = 160$ s at least, the original solution with $N = 100$ is accurate to the resolution of the graph.

To assess the differences at $t = 160$ s more carefully, we might invoke OCTAVE to calculate the difference, say, between the solution at that time for $N = 100$ and the solution for $N = 200$. To do so, we would invoke the statements\footnote{Remember that the solution for $N = 200$ has twice as many values at the solution for $N = 100$.}

\begin{verbatim}
    diff = zeros(101,1);
    for i=1:101 diff(i)=u1_200(2*i-1)-u1_100(i); end
    max( diff )
    ans = 2.2048e-004
\end{verbatim}

The two solutions are extremely close throughout the rod. We conclude that the solution obtained for $N = 100$ is probably accurate to an absolute error of about $\pm 2 \times 10^{-4}$.

15.19.3 The Laplace Equation

The final step in addressing the example laid out in Section 15.16.3 is to set up and solve Eq. (15.174). Using a command file in OCTAVE, we would begin by requesting input of the two necessary parameters and assuring that each is stored with the proper data type. The statements

\begin{verbatim}
plot( x_400, u1_400, 'linewidth', 3, 'color', 'black' )
title( 'N = 400', 'fontsize', 14 )
set(gca, 'YLim', [0.0,0.6] )
\end{verbatim}
15.19. USING OCTAVE TO SOLVE 2D PDES VIA AN FDM

\[ N = \text{input('Enter number of segments (N): ')}; \ N = \text{fix}(N); \]
\[ L = \text{input('Enter length of side(L): ')}; \]

accomplish those objectives. Then, we would determine the length of each segment and establish values for the coordinates at which a solution will be generated with the statements

\[
\text{dx} = L/N; \\
\text{x} = \text{dx}*[ \ 0 : \ N ]; \ y = \text{x};
\]

(Remember that we are solving the Laplace equation in a square and we have elected to use the same segment size for both coordinates.)

Next, we establish a suitably dimensioned array \( u \) to contain the solution as it is generated, set the Dirichlet boundary values on the left, bottom, and top edges, and display the initial values in the array \( u \) with the statements

\[ u:\text{zeros}(N+1, N+1); \]
\[ u(1,:) = 100.0; \]
\[ u(:,1) = 100.0 - 100.0*x/L; \]
\[ u \]

In constructing these statements, we have arranged the values in the array \( u \) so that, in the output generated by the statement \textit{print}, \( u \), the left, right, top, and bottom edges correspond to the left, right, top, and bottom edges in Fig. 15.13. Thus, for example, execution of the statements introduced to this point with \( N \) set to 5 and \( L \) set to 10 yields the output

\[
\begin{array}{cccccccc}
100 & 100 & 100 & 100 & 100 & 100 \\
80 & 0 & 0 & 0 & 0 & 0 \\
60 & 0 & 0 & 0 & 0 & 0 \\
40 & 0 & 0 & 0 & 0 & 0 \\
20 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
\end{array}
\]

The values in the top and bottom rows and in the left column of this array correctly reflect the Dirichlet conditions to be imposed on the solution. The values in the right column are unknown initially, since we have imposed Neumann (derivative) conditions along that edge. Those values must be determined as the solution unfolds.

Our task from this point is to determine values to assign to the entries not on the top, bottom, or left edges in such a way that Eq. (15.174) is satisfied. This task differs substantially from the task we confronted in solving the wave and diffusion equations in Sections 15.19.1 and 15.19.2. Those previous equations involved a mixture of boundary and initial conditions and, once the process was started, we could step forward in time as far as we pleased. With the Laplace equation, we are dealing with a boundary value problem alone, and conditions are imposed around the \textit{entire} perimeter of the region of interest. Rather than having boundary conditions on the left and right of the region of interest, initial conditions on the bottom, and no definite boundary on the top, we have instead boundary conditions on all four sides of the region of interest. One strategy for the present problem would be to \textit{guess} the derivative \( \partial u/\partial x \) at the left edge, treat the problem as an initial value problem in \( x \) by stepping systematically across the above array to the right edge, and hope that we arrived at that edge with values satisfying the boundary condition at that edge. If we missed, we would make another guess for the starting derivative and try again. We conclude that, in contrast to the problems treated in Sections 15.19.1 and 15.19.2, solution of this problem will require an \textit{iterative} approach.
We have already written Eq. (15.174) to support an iterative approach more suitable than that suggested in the previous paragraph. This approach entails (1) guessing a solution at each node not constrained by a Dirichlet condition, (2) stepping systematically through those nodes while replacing the value at each node with the value determined by Eq. (15.174), and (3) repeating step (2) until some criterion of convergence has been met. For the moment, let us simply carry out a user-specified number of iterations. Thus, we would add to the read statements the coding

```matlab
maxits = input( 'Maximum number of iterations (maxits): ');
maxits = fix( maxits );
```

to obtain the desired number of iterations, we would change the `print` statement above to read

```matlab
printf( '\nIteration 0\n' )
u
```

so as to label the output and space it away from what precedes and follows and, finally, we would add the multiple loop

```matlab
for itcnt=1:maxits
    for i = 2:N
        for j = 2:N
            u(i,j) = 0.25*(u(i+1,j) + u(i-1,j) + u(i,j+1) + u(i,j-1));
        end
        u(i,N+1) = 0.25*(2.0*u(i,N) + u(i-1,N+1) + u(i+1,N+1));
    end
    printf( '\nIteration %d\n', itcnt )
    u
end
```

to effect the solution. Here, each pass through the outermost loop will effect one iteration in the algorithm described above. A total of `maxits` iterations will be executed before the loop terminates. Within that outermost loop, the double loop on `i` and `j` steps through the “interior” points in the grid, replacing the value at each with the average of the values at its four nearest neighbors. Then, outside the innermost loop (the loop on `j`), the one statement that remains in the loop on `i` through the nodes along the right edge (except for the two corner nodes), replacing the value at each with the value dictated by the appropriate member of Eq. (15.174). Finally, the `printf` statement labels the iteration and the single statement `u` displays the solution after the current iteration is completed. A more fully commented command file containing these statements—and a few others to be discussed in a moment—is named `fdmlap2d.m`, is listed in Appendix ??, and can be copied from the directory `$HEAD/octave`.

As a quick test of our as yet incomplete command file and as a way to illustrate the manner of approach to a final solution, let us execute the program as it stands with the input

```matlab
fdmlap2d
Enter number of segments (N): 5
Enter length of side(L): 10
Maximum number of iterations: 5
```

40. This iterative approach has already been described—though for slightly different boundary conditions—in Section 9.3.1 and implemented in OCTAVE in Section 9.10.

41. We elect to work in place in the array that contains our initial guess, so each newly generated value is used in subsequent calculations as soon as it becomes available. Alternatively, we could store the newly generated values in a second array, completing one pass through the array before using any of the newly generated values. As it turns out, the former approach converges rather more rapidly than the latter approach.
15.19. USING OCTAVE TO SOLVE 2D PDES VIA AN FDM

obtaining the output

<table>
<thead>
<tr>
<th>Iteration 0</th>
<th>u = 100 100 100 100 100 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>80 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>60 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>40 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>20 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>0 0 0 0 0 0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Iteration 1</th>
<th>u = 100.0000 100.0000 100.0000 100.0000 100.0000 100.0000</th>
</tr>
</thead>
<tbody>
<tr>
<td>80.0000 45.0000 36.2500 34.0625 33.5156 41.7578</td>
<td></td>
</tr>
<tr>
<td>60.0000 26.2500 15.6250 12.4219 11.4844 16.1816</td>
<td></td>
</tr>
<tr>
<td>40.0000 16.5625 8.0469 5.1172 4.1504 6.1207</td>
<td></td>
</tr>
<tr>
<td>20.0000 9.1406 4.2969 2.3535 1.6259 2.3431</td>
<td></td>
</tr>
<tr>
<td>0.0000 0.0000 0.0000 0.0000 0.0000 0.0000</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Iteration 2</th>
<th>u = 100.0000 100.0000 100.0000 100.0000 100.0000 100.0000</th>
</tr>
</thead>
<tbody>
<tr>
<td>80.0000 60.6250 52.5781 49.6289 50.7178 54.0430</td>
<td></td>
</tr>
<tr>
<td>60.0000 38.2031 27.8125 23.5107 23.6401 26.9513</td>
<td></td>
</tr>
<tr>
<td>40.0000 23.8477 15.2686 11.3208 10.6769 12.6205</td>
<td></td>
</tr>
<tr>
<td>20.0000 12.0361 7.4146 5.0903 4.5276 5.4293</td>
<td></td>
</tr>
<tr>
<td>0.0000 0.0000 0.0000 0.0000 0.0000 0.0000</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Iteration 3</th>
<th>u = 100.0000 100.0000 100.0000 100.0000 100.0000 100.0000</th>
</tr>
</thead>
<tbody>
<tr>
<td>80.0000 67.6953 61.2841 58.8781 59.2306 61.3531</td>
<td></td>
</tr>
<tr>
<td>60.0000 44.8389 36.2256 32.5162 32.3438 34.6757</td>
<td></td>
</tr>
<tr>
<td>40.0000 28.0359 20.7492 17.2582 16.6978 18.3751</td>
<td></td>
</tr>
<tr>
<td>0.0000 0.0000 0.0000 0.0000 0.0000 0.0000</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Iteration 4</th>
<th>u = 100.0000 100.0000 100.0000 100.0000 100.0000 100.0000</th>
</tr>
</thead>
<tbody>
<tr>
<td>80.0000 71.5308 66.6586 64.6013 64.5746 65.9620</td>
<td></td>
</tr>
<tr>
<td>60.0000 48.9481 42.2180 39.1053 38.7634 40.4643</td>
<td></td>
</tr>
<tr>
<td>40.0000 30.8899 25.0729 22.2009 21.7133 23.0604</td>
<td></td>
</tr>
<tr>
<td>20.0000 15.2038 12.0512 10.4417 10.1263 10.8283</td>
<td></td>
</tr>
<tr>
<td>0.0000 0.0000 0.0000 0.0000 0.0000 0.0000</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Iteration 5</th>
<th>u = 100.0000 100.0000 100.0000 100.0000 100.0000 100.0000</th>
</tr>
</thead>
<tbody>
<tr>
<td>80.0000 73.9017 70.1803 68.4650 68.2961 69.2642</td>
<td></td>
</tr>
<tr>
<td>60.0000 51.7524 46.5277 43.9892 43.6185 44.8890</td>
<td></td>
</tr>
<tr>
<td>40.0000 33.0073 28.4467 26.1477 25.7375 26.7981</td>
<td></td>
</tr>
<tr>
<td>0.0000 0.0000 0.0000 0.0000 0.0000 0.0000</td>
<td></td>
</tr>
</tbody>
</table>

The progression in the solution from iterate to iterate is quite clear, though it is also clear from the changes taking place from one iterate to the next that we haven’t gone far enough to achieve
convergence.

While the coding developed to this point certainly displays the essence of the algorithm for solving the Laplace equation, the above output also makes clear that we must carry the solution through a larger number of iterates to achieve convergence. That change, however, will result in much more output unless we suppress all but every \( f \)-th iteration by adding the coding

\[
\text{f = input('Display frequency (f): ');} \; \text{f} = \text{fix(f)};
\]

to request the number of iterates whose display is to be suppressed and, in addition, by modifying the print statement in the final loops to make the display conditional upon \( \text{itcnt} \) being a multiple of \( f \). In effect, we simply replace all but the first \textit{printf} statement everywhere it appears with the conditional statement

\[
\text{if f*fix(itcnt/f) == itcnt}
\text{printf('Iteration \%d', itcnt)}
\text{u}
\text{end}
\]

(See the full listing in Appendix 15.D.3 for the details.)

With this embellishment, we might test our developing program by submitting the starting values

```
fdmlap2d
Number of segments: 5
Length of side: 10
Maximum number of iterations: 80
Display frequency: 20
```

obtaining the output

**Iteration 0**

\[
\begin{array}{cccccc}
\text{u} & 100 & 100 & 100 & 100 & 100 \\
80 & 0 & 0 & 0 & 0 & 0 \\
60 & 0 & 0 & 0 & 0 & 0 \\
40 & 0 & 0 & 0 & 0 & 0 \\
20 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
\end{array}
\]

**Iteration 20**

\[
\begin{array}{cccccccc}
\text{u} & 100.0000 & 100.0000 & 100.0000 & 100.0000 & 100.0000 & 100.0000 \\
80.0000 & 79.8816 & 79.8018 & 79.7599 & 79.7516 & 79.7516 & 79.7516 \\
60.0000 & 59.8314 & 59.7178 & 59.6582 & 59.6464 & 59.6464 & 59.6464 \\
0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\
\end{array}
\]

**Iteration 40**

\[
\begin{array}{cccccccc}
\text{u} & 100.0000 & 100.0000 & 100.0000 & 100.0000 & 100.0000 & 100.0000 \\
80.0000 & 79.9992 & 79.9988 & 79.9985 & 79.9985 & 79.9985 & 79.9985 \\
60.0000 & 59.9989 & 59.9983 & 59.9979 & 59.9979 & 59.9979 & 59.9979 \\
\end{array}
\]
15.19. USING OCTAVE TO SOLVE 2D PDES VIA AN FDM

Iteration 60

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>80.00000</td>
<td>80.00000</td>
<td>79.99999</td>
<td>79.99999</td>
<td>79.99999</td>
<td>79.99999</td>
</tr>
<tr>
<td>60.00000</td>
<td>59.99999</td>
<td>59.99999</td>
<td>59.99999</td>
<td>59.99999</td>
<td>59.99999</td>
</tr>
<tr>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
</tbody>
</table>

Iteration 80

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>100.0000</td>
<td>100.0000</td>
<td>100.0000</td>
<td>100.0000</td>
<td>100.0000</td>
<td>100.0000</td>
</tr>
<tr>
<td>80.00000</td>
<td>80.00000</td>
<td>80.00000</td>
<td>80.00000</td>
<td>80.00000</td>
<td>80.00000</td>
</tr>
<tr>
<td>60.00000</td>
<td>60.00000</td>
<td>60.00000</td>
<td>60.00000</td>
<td>60.00000</td>
<td>60.00000</td>
</tr>
<tr>
<td>40.00000</td>
<td>40.00000</td>
<td>40.00000</td>
<td>40.00000</td>
<td>40.00000</td>
<td>40.00000</td>
</tr>
<tr>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
</tbody>
</table>

In this output, convergence to a steady solution is more evident. Indeed, the solution at iteration 80 conforms to exactly what we might expect.

While the coarse grid used for the above solution renders its display easy on the page of a book and, in particular, allows us to reveal the nature of the iterative convergence to the final steady-state solution, that coarse grid also leaves open the possibility that we have found an accurate solution to the discretized equations that is not, however, a particularly good approximation to the continuous solution to the original problem.\textsuperscript{42} We would prefer to use a more refined grid and to use graphical rather than tabular display of the solution. To achieve those objectives, we need simply specify a larger number of segments, a larger number of iterations, a larger gap between displayed iterates, and the plotting of a mesh surface with the statement\textsuperscript{43} 

$$\text{mesh}(x,y,\text{flip}(u,1), 'edgecolor', 'black')$$

as an alternative to the printing of a table on the screen with \texttt{printf}. A more fully commented command file containing these modifications—and a few others to be discussed in a moment—is named \texttt{fdmlap2d_plot.m}, is listed in Appendix 15.D.4, and can be copied from the directory \texttt{$\$HEAD/octave}.

With the changes described in the previous paragraph, the solution is generated and displayed so quickly that we cannot easily follow its evolution. We fix this awkwardness by introducing a time delay after each plot is produced by inserting the statement

$$\text{pause( delay )}$$

after each \texttt{mesh} statement and inserting at the beginning the statement

$$\text{delay = input( 'Delay between plots in secs (delay): ' )};$$

so the user has control over that delay at execution time and so that OCTAVE is configured to respond to the \texttt{pause} statement.

\textsuperscript{42}In fact, the solution we have obtained is exactly correct, but the problem we have solved is also especially simple. Actually obtaining the exact solution with such a coarse grid is much more an accident of the particular problem than a behavior we can expect.

\textsuperscript{43}The command \texttt{flip} is used so that the orientation of the solution on the axes in the resulting graphs is the same as the orientation of the printed arrays earlier in this section.
In a first execution of this coding, however, the graphical displays are oriented so as to tempt the eye into viewing the surface from the wrong side. To repair that problem, we exploit the rotation of the surface in the plot window to find a reasonable orientation, execute the statement `[a,e]=view()` to find the azimuth and elevation associated with that view, and add the statements

```
xlabel('x','fontsize',16); ylabel('y','fontsize',16);
zlabel('u','fontsize',16);
set(gca,'fontsize',12)
view([40 10])
```

immediately after each occurrence of the `mesh` command to adjust the size of the labels on and the default orientation of the graph. (See the listing in Appendix 15.D.4 for the details of this change.)

Once all of these changes have been implemented, we generate an animation showing the convergence of the iterates towards a final solution with the statements

```
fdmlap2d_plot
```

```
Enter number of segments (N): 20
Enter length of side(L): 10
Maximum number of iterations (maxits): 80
Display frequency (f): 10
Delay between plots in secs (delay): .5
```

The output of this program for several steps along the way from the initial guess to the final equilibrium is shown in Fig. 15.18. The upper left panel in this figure shows the initial guess, to the resolution specified by division of each edge into 20 segments. Subsequent panels are separated by 25 iterations. The gradual penetration of thermal energy from the edge along which the temperature is fixed and varies linearly towards the parallel insulated edge is clear as the iterations proceed. The front and back edges also remain at the steady temperature defined by the boundary conditions. While 75 iterations are clearly not enough for the solution to relax completely, we nonetheless anticipate—correctly—that the ultimate equilibrium solution will appear in this display as a plane whose front edge is at 0 °C and whose back edge is at 100 °C.

In the above discussion, we have simply allowed some specified number of iterations to be carried out and applied qualitative reality checks to the resulting solutions. A common means to assess convergence of the iterative solution of the discretized equations involves comparing the values generated in each iterate with the ones from the previous iterate and accepting the most recent iterate as final once it differs from its predecessor by no more than a specified tolerance. One then supposes that the iterative algorithm has converged to an accuracy—fractional or absolute depending on the precise coding—given by the tolerance. Modification of `fdm2dlap_plot.m` to incorporate this embellishment is the subject of Exercise 15.23.

Convergence of the iterative process to a particular tolerance, however, only provides evidence that the discretized equations have been solved to a particular accuracy. That characteristic assures us only that convergence error is adequately small. (Probably roundoff error is also under control at that point.) By itself, convergence of the iterative process offers only marginal assurance that discretization error is adequately small. To assess discretization error, we would need to compare not successive iterates on a given grid but converged iterates on at least two grids, one of which is finer than the other. Multigrid algorithms to achieve that objective are beyond the scope of this chapter.

15.23 Finite Element Methods (FEMs) in Two Dimensions

While their implementation is more involved in two dimensions than in one dimension, finite element approaches to problems in two dimensions involve the same several steps as those identified at the
Figure 15.18: Approach of temperature distribution to equilibrium. The panels show the initial distribution (upper left), the distribution at iterate 25 (upper right), the distribution at iterate 50 (lower left), and the distribution at iterate 75 (lower right). These figures are oriented so that the previously displayed arrays can simply be laid down on the $xy$ plane in the figures. This graph was produced with OCTAVE.

beginning of Section 15.9. In this section, we illustrate the method of finite element analysis by applying those steps to the two-dimensional boundary value problem defined in Section 15.1.8.

15.23.1 Discretizing the Domain: Pre-processing

The first step is to divide the domain into elements or, in the jargon, to create a mesh over the domain or to mesh the domain. To facilitate the creation of that mesh (and the discussion of the method), we elect to use the simplest possible two-dimensional element, meshing the region of interest—denote it by $\Omega$—with triangular elements as illustrated, for example, in the left panel of Fig. 15.19. Each element is defined by three nodes (vertices), which also serve as nodes on adjacent elements, and each element has three edges—the lines along which pairs of elements meet. Together, the edges define the perimeter or boundary of the element. For the meshing to be legal, no node in one element can be located along the edge of another element, as, for example, at the point circled in the right panel of Fig. 15.19. Further, since the error in the finite element solution varies inversely as the sine of the smallest angle in each element, all elements should be at least approximately equilateral. While increasing the number of elements in the region improves the accuracy of the solution, it also increases both the computation time and the memory needed to obtain the solution. Thus, the number of elements should be large (and their size small) only in regions in which the
solution is expected to vary rapidly. A coarser mesh can be accepted in regions in which the solution varies slowly. Choosing an appropriate mesh of elements of suitable sizes is difficult and may require an iterative approach, especially if little is known \textit{a priori} about the solution.

As a modest example, suppose the region of interest is a rectangle. For simplicity, divide that region into eight triangular elements by one vertical line, one horizontal line, and three diagonal lines, as shown in Fig. 15.20. Then, we number the elements in an order that seems convenient. In addition, we number the nodes both globally (i.e., with respect to the entire domain) and locally (i.e., with respect to a particular element). The \textit{global} labeling is done by assigning a specific and unique integer to each node, as in the left panel of Fig. 15.21. \textit{Locally}, the nodes—remember that each element has three nodes—are numbered with the integers 0, 1, and 2, as in the right panel of Fig. 15.21.\footnote{In this example, we elect to begin all numbering of elements and nodes—both global and local—with zero, fully aware that some programming languages (IDL, PYTHON, C) follow that pattern but that others (MATLAB, OCTAVE, FORTRAN) start numbering with 1. In the latter case, each index in the program will be incremented by 1 relative to those in this theoretical discussion of the general strategies.} Each element will have its own zeroth, first, and second nodes, each of which also has a global label that identifies it uniquely. A connectivity matrix such as the one shown in Table 15.2 is created to express the relation between global and local node numbers. In the present context, the matrix consists of $M$ rows, one for each element, with each row containing 3—one for each node
Figure 15.21: Rectangular domain of Fig. 15.20 with global (left) and local (right) node numbers.

Table 15.2: The connectivity matrix for the example in Section 15.23.

<table>
<thead>
<tr>
<th>Element</th>
<th>Global number of local node 0</th>
<th>Global number of local node 1</th>
<th>Global number of local node 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>8</td>
<td>7</td>
</tr>
</tbody>
</table>

of the element—entries, i.e. the matrix has $M$ rows and three columns. Every entry in the matrix is a global node number corresponding to the zeroth, first, or second node of each element. In this example, the (global) numbering both of the elements and of the nodes is done in the same way, from top to bottom and from left to right. The local node numbers are assigned counterclockwise starting at the lower left node of each element. For element 0, for example, the global number of local node 0 is 1, the global number of local node 1 is 3, and the global number of local node 2 is 0. Similarly, for element 1, we have global node numbers 1, 4, and 3. The rest of the entries are generated in the same way and are recorded in Table 15.2.

### 15.23.2 Selecting the Interpolation or Shape Functions

The simplest interpolation or shape functions are linear in both $x$ and $y$. For element $e$, we write the approximating equation in the form

$$\tilde{\varphi}^{(e)}(x, y) = a^{(e)} + b^{(e)}x + c^{(e)}y$$

(15.178)
The constants are determined so that this equation gives the correct values at all three nodes of any element, i.e., so that

\[ \tilde{\varphi}^{(e)}_0 = a^{(e)} + b^{(e)} x_0^{(e)} + c^{(e)} y_0^{(e)} \]
\[ \tilde{\varphi}^{(e)}_1 = a^{(e)} + b^{(e)} x_1^{(e)} + c^{(e)} y_1^{(e)} \]  
\[ \tilde{\varphi}^{(e)}_2 = a^{(e)} + b^{(e)} x_2^{(e)} + c^{(e)} y_2^{(e)} \]  

(15.179)

where subscripts refer to local node numbers. Solution of this system of equations for \( a^{(e)} \), \( b^{(e)} \), and \( c^{(e)} \) and substitution of the solution into Eq. (15.178) gives the approximating function

\[ \tilde{\varphi}^{(e)}(x, y) = \sum_{j=0}^{2} N_j^{(e)}(x, y) \tilde{\varphi}^{(e)}_j \]  

(15.180)

where, with \( \Delta^{(e)} \)—the determinant of the matrix of coefficients in Eq. (15.179)—defined by

\[ \Delta^{(e)} = \begin{vmatrix} 1 & x_0^{(e)} & y_0^{(e)} \\ 1 & x_1^{(e)} & y_1^{(e)} \\ 1 & x_2^{(e)} & y_2^{(e)} \end{vmatrix} \]  

(15.181)

the interpolation or shape functions \( N_j^{(e)}(x, y) \) are given by

\[ N_0^{(e)}(x, y) = \frac{1}{\Delta^{(e)}} \begin{vmatrix} 1 & x & y \\ 1 & x_1^{(e)} & y_1^{(e)} \\ 1 & x_2^{(e)} & y_2^{(e)} \end{vmatrix} \]  

(15.182)

\[ N_1^{(e)}(x, y) = \frac{1}{\Delta^{(e)}} \begin{vmatrix} 1 & x & y \\ 1 & x_0^{(e)} & y_0^{(e)} \\ 1 & x_2^{(e)} & y_2^{(e)} \end{vmatrix} \]  

(15.183)

\[ N_2^{(e)}(x, y) = \frac{1}{\Delta^{(e)}} \begin{vmatrix} 1 & x & y \\ 1 & x_0^{(e)} & y_0^{(e)} \\ 1 & x_1^{(e)} & y_1^{(e)} \end{vmatrix} \]  

(15.184)

From these expressions and the properties of determinants, we can quickly recognize that each interpolation function has the value one at the node corresponding to its index and the value zero at the other two nodes. Thus, the sum on the right side of Eq. (15.180) reduces at each node to the value given on the left side. Equally true but less obvious, each interpolation function is in fact zero along the entire line joining the two nodes at which it has the value zero. This property will turn out to be useful when the boundary conditions are taken into account.

15.23.3 Formulating the Equations for a Single Element

From the general form of Eq. (15.70), the expression for the residual \( r \) can be written easily as

\[ r = -\frac{\partial}{\partial x} \left( \alpha_x \frac{\partial \tilde{\varphi}}{\partial x} \right) - \frac{\partial}{\partial y} \left( \alpha_y \frac{\partial \tilde{\varphi}}{\partial y} \right) + \beta \tilde{\varphi} - f \]  

(15.185)

\[45\text{See Exercise 15.15.}\]
\[46\text{See Exercise 15.15.}\]
Evaluated for element $e$, the weighted residuals then are given by the integral

$$ R_i^{(e)} = \int \int_{\Omega^{(e)}} N_i^{(e)} r \, dx \, dy; \quad i = 0, 1, 2 $$

(15.186)

over the surface $\Omega^{(e)}$ of element $e$. Here, following the Galerkin formulation, we take the weighting functions to be the interpolation functions themselves. Substitution of Eq. (15.185) into Eq. (15.186) leads to

$$ R_i^{(e)} = \int \int_{\Omega^{(e)}} N_i^{(e)} \left[ -\frac{\partial}{\partial x} \left( \alpha_x \frac{\partial \tilde{\phi}}{\partial x} \right) - \frac{\partial}{\partial y} \left( \alpha_y \frac{\partial \tilde{\phi}}{\partial y} \right) + \beta \tilde{\phi} - f \right] \, dx \, dy $$

(15.187)

$$ = -\int \int_{\Omega^{(e)}} N_i^{(e)} \left[ \frac{\partial}{\partial x} \left( \alpha_x \frac{\partial \tilde{\phi}}{\partial x} \right) + \frac{\partial}{\partial y} \left( \alpha_y \frac{\partial \tilde{\phi}}{\partial y} \right) \right] \, dx \, dy $$

$$ + \int \int_{\Omega^{(e)}} N_i^{(e)} \beta \tilde{\phi} \, dx \, dy - \int \int_{\Omega^{(e)}} N_i^{(e)} f \, dx \, dy $$

(15.188)

Then, interpreting $\psi$, $V_x$, and $V_y$ in Eq. (15.232), the identity developed in Appendix 15.E, as $N_i^{(e)}$, $\alpha_x \partial \tilde{\phi}/\partial x$, and $\alpha_y \partial \tilde{\phi}/\partial y$, respectively, we can rewrite the first integral in Eq. (15.188) so as to recast the entirety of Eq. (15.188) in the form

$$ R_i^{(e)} = \int \int_{\Omega^{(e)}} \left( \alpha_x \frac{\partial N_i^{(e)} \partial \tilde{\phi}}{\partial x} + \alpha_y \frac{\partial N_i^{(e)} \partial \tilde{\phi}}{\partial y} + \beta N_i^{(e)} \tilde{\phi} \right) \, dx \, dy $$

$$ - \int \int_{\Omega^{(e)}} N_i^{(e)} f \, dx \, dy - \oint_{\Gamma^{(e)}} N_i^{(e)} \left( \alpha_x \frac{\partial \tilde{\phi}}{\partial x} \hat{i} + \alpha_y \frac{\partial \tilde{\phi}}{\partial y} \hat{j} \right) \cdot \hat{n}^{(e)} \, dl $$

(15.189)

Here, $\Gamma^{(e)}$ is the path bounding the region $\Omega^{(e)}$, and $\hat{n}^{(e)}$ is a unit vector lying in the plane of element $e$, perpendicular at each point to $\Gamma^{(e)}$, and directed outward from the perspective of a viewer in $\Omega^{(e)}$. Substitution of the approximating function of Eq. (15.180) yields for the weighted residuals the expression

$$ R_i^{(e)} = \sum_{j=0}^{2} \varphi_j^{(e)} \int \int_{\Omega^{(e)}} \left( \alpha_x \frac{\partial N_j^{(e)} \partial \tilde{\phi}}{\partial x} + \alpha_y \frac{\partial N_j^{(e)} \partial \tilde{\phi}}{\partial y} + \beta N_j^{(e)} \tilde{\phi} \right) \, dx \, dy $$

$$ - \int \int_{\Omega^{(e)}} N_i^{(e)} f \, dx \, dy - \oint_{\Gamma^{(e)}} N_i^{(e)} \left( \alpha_x \frac{\partial \tilde{\phi}}{\partial x} \hat{i} + \alpha_y \frac{\partial \tilde{\phi}}{\partial y} \hat{j} \right) \cdot \hat{n}^{(e)} \, dl $$

(15.190)

or, in the matrix form,

$$ \{R^{(e)}\} = [K^{(e)}]\{\varphi^{(e)}\} - \{b^{(e)}\} - \{g^{(e)}\} $$

(15.191)

where $\ldots$ denotes a matrix and $\{\ldots\}$ denotes a vector. The matrix $[K^{(e)}]$ is a $3 \times 3$ matrix and $\{\varphi^{(e)}\}$, $\{b^{(e)}\}$, and $\{g^{(e)}\}$ are three-element vectors. The matrix elements $K_{ij}^{(e)}$ and the vector elements $b_i^{(e)}$ and $g_i^{(e)}$ can be obtained from Eq. (15.190), namely

$$ K_{ij}^{(e)} = \int \int_{\Omega^{(e)}} \left( \alpha_x \frac{\partial N_j^{(e)} \partial \tilde{\phi}}{\partial x} + \alpha_y \frac{\partial N_j^{(e)} \partial \tilde{\phi}}{\partial y} + \beta N_j^{(e)} N_j^{(e)} \right) \, dx \, dy $$

(15.192)

$$ b_i^{(e)} = \int \int_{\Omega^{(e)}} N_i^{(e)} f \, dx \, dy $$

(15.193)

$$ g_i^{(e)} = \oint_{\Gamma^{(e)}} N_i^{(e)} \left( \alpha_x \frac{\partial \tilde{\phi}}{\partial x} \hat{i} + \alpha_y \frac{\partial \tilde{\phi}}{\partial y} \hat{j} \right) \cdot \hat{n}^{(e)} \, dl $$

(15.194)

where, in all cases, $i$ and $j$ assume the values 0, 1, and 2.
All elements of the matrix \([K^{(e)}]\) and of the vector \(\{b^{(e)}\}\) are known at the start of the problem. The elements of the vector \(\{g^{(e)}\}\) are ultimately determined when the boundary conditions are incorporated. Those terms will be discussed further in Section 15.23.5.

Of course, the equations determining the unknown nodal values are obtained by requiring the weighted residuals to be zero. In matrix form, we have from Eq. (15.191) for the \(e\)-th element that

\[
[K^{(e)}]\{\varphi^{(e)}\} = \{b^{(e)}\} + \{g^{(e)}\} \tag{15.195}
\]

Written out in terms of the specific (local) nodes, this equation becomes

\[
\begin{bmatrix}
K^{(e)}_{00} & K^{(e)}_{01} & K^{(e)}_{02} \\
K^{(e)}_{10} & K^{(e)}_{11} & K^{(e)}_{12} \\
K^{(e)}_{20} & K^{(e)}_{21} & K^{(e)}_{22}
\end{bmatrix}
\begin{bmatrix}
\varphi^{(e)}_0 \\
\varphi^{(e)}_1 \\
\varphi^{(e)}_2
\end{bmatrix}
= 
\begin{bmatrix}
b^{(e)}_0 \\
b^{(e)}_1 \\
b^{(e)}_2
\end{bmatrix}
+ 
\begin{bmatrix}
g^{(e)}_0 \\
g^{(e)}_1 \\
g^{(e)}_2
\end{bmatrix} \tag{15.196}
\]

15.23.4 Assembling the System of Equations

The goal of assembly is to generate a system of equations that can be solved for the vector \(\{\varphi\}\), whose—here 9—components approximate the solution to the original problem at the nodes. That is, we seek a set of equations of the form

\[
[K]\{\varphi\} = \{b\} + \{g\} \tag{15.197}
\]

where \([K]\) is a—here 9 \times 9—matrix and \(\{b\}\) and \(\{g\}\) are—here 9 component—vectors. As in the one-dimensional case discussed in Section 15.9, the equations for the individual elements are assembled to take into account the requirement that the solution be continuous at the boundaries between elements. Regardless of which element one focuses on, the solution along each bounding line is a linear interpolation between the values at the two nodes defining the line. Thus, requiring that the solution be continuous at each node assures continuity along the lines joining all nodes. To illustrate the process, consider the assembly of the equations relating to elements 0 and 1. As shown in Fig. 15.21 and in the connectivity matrix, the \textit{global} numbers of the three nodes of element 0 are (in order) 1, 3, and 0. Thus,

\[
\varphi^{(0)}_0 = \varphi_1 ; \quad \varphi^{(0)}_1 = \varphi_3 ; \quad \varphi^{(0)}_2 = \varphi_0 \tag{15.198}
\]

and, written with \textit{global} identifications on the unknown \(\varphi\)'s, Eq. (15.196) for element 0 becomes

\[
\begin{bmatrix}
K^{(0)}_{00} & K^{(0)}_{01} & K^{(0)}_{02} \\
K^{(0)}_{10} & K^{(0)}_{11} & K^{(0)}_{12} \\
K^{(0)}_{20} & K^{(0)}_{21} & K^{(0)}_{22}
\end{bmatrix}
\begin{bmatrix}
\varphi_1 \\
\varphi_3 \\
\varphi_0
\end{bmatrix}
= 
\begin{bmatrix}
b^{(0)}_0 \\
b^{(0)}_1 \\
b^{(0)}_2
\end{bmatrix}
+ 
\begin{bmatrix}
g^{(0)}_0 \\
g^{(0)}_1 \\
g^{(0)}_2
\end{bmatrix} \tag{15.199}
\]

Similarly, the three nodes of element 1 are (in order) 1, 4, and 3. Thus,

\[
\varphi^{(1)}_0 = \varphi_1 ; \quad \varphi^{(1)}_1 = \varphi_4 ; \quad \varphi^{(1)}_2 = \varphi_3 \tag{15.200}
\]

and Eq. (15.196) for element 1 becomes

\[
\begin{bmatrix}
K^{(1)}_{00} & K^{(1)}_{01} & K^{(1)}_{02} \\
K^{(1)}_{10} & K^{(1)}_{11} & K^{(1)}_{12} \\
K^{(1)}_{20} & K^{(1)}_{21} & K^{(1)}_{22}
\end{bmatrix}
\begin{bmatrix}
\varphi_1 \\
\varphi_4 \\
\varphi_3
\end{bmatrix}
= 
\begin{bmatrix}
b^{(1)}_0 \\
b^{(1)}_1 \\
b^{(1)}_2
\end{bmatrix}
+ 
\begin{bmatrix}
g^{(1)}_0 \\
g^{(1)}_1 \\
g^{(1)}_2
\end{bmatrix} \tag{15.201}
\]

To illustrate the process, consider the assembly of the equations relating to elements 0 and 1. As shown in Fig. 15.21 and in the connectivity matrix, the \textit{global} numbers of the three nodes of element 0 are (in order) 1, 3, and 0. Thus,
We have, of course, found six equations constraining only four unknowns, namely $\tilde{\varphi}_0$, $\tilde{\varphi}_1$, $\tilde{\varphi}_3$, and $\tilde{\varphi}_4$. Two of the six are redundant. Rather than discard two of them, however, we elect to reduce the number to four by adding selected pairs of these equations. To see the proper combinations, let us recast each of Eqs. (15.199) and (15.201) as equations for the vector containing all four of the involved nodal values. That is, let us augment each equation to make it more obvious that it is not the only equation with which we must deal. In the process, we add, for example, to Eq. (15.199) a row and a column corresponding to global node 4 even though that node does not enter into the equations for element 0. Further, we rearrange the order of the columns in the matrix of Eq. (15.199) so the vector of unknowns can be written with the entries in the order of the global node numbers and we rearrange the order of the rows so that the entries in $\{b\}$ and $\{g\}$ are also in the order of the global node numbers (and, incidentally and automatically, so that the symmetry of the augmented matrix is preserved). The result is

$$
\begin{bmatrix}
K_{22}^{(0)} & K_{20}^{(0)} & K_{21}^{(0)} \\
K_{02}^{(0)} & K_{00}^{(0)} & K_{01}^{(0)} \\
K_{12}^{(0)} & K_{10}^{(0)} & K_{11}^{(0)} \\
0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
\tilde{\varphi}_0 \\
\tilde{\varphi}_1 \\
\tilde{\varphi}_3 \\
\tilde{\varphi}_4
\end{bmatrix}
= 
\begin{bmatrix}
b_{22}^{(0)} \\
b_{02}^{(0)} \\
b_{12}^{(0)} \\
0
\end{bmatrix}
+ 
\begin{bmatrix}
g_2^{(0)} \\
g_0^{(0)} \\
g_1^{(0)} \\
0
\end{bmatrix}
$$

(15.202)

Similarly, we augment Eq. (15.201) to obtain the result

$$
\begin{bmatrix}
0 & 0 & 0 & 0 \\
0 & K_{20}^{(1)} & K_{22}^{(1)} & K_{21}^{(1)} \\
0 & K_{10}^{(1)} & K_{12}^{(1)} & K_{11}^{(1)}
\end{bmatrix}
\begin{bmatrix}
\tilde{\varphi}_0 \\
\tilde{\varphi}_1 \\
\tilde{\varphi}_3 \\
\tilde{\varphi}_4
\end{bmatrix}
= 
\begin{bmatrix}
0 \\
b_{02}^{(1)} \\
b_{12}^{(1)} \\
0
\end{bmatrix}
+ 
\begin{bmatrix}
g_0^{(1)} \\
g_1^{(1)}
\end{bmatrix}
$$

(15.203)

Basically, we reduce the original set of six equations to the required four by adding these equations, finding as the result of assembling the equations for elements 0 and 1 that

$$
\begin{bmatrix}
K_{22}^{(0)} & K_{20}^{(0)} & K_{21}^{(0)} \\
K_{02}^{(0)} & K_{00}^{(0)} + K_{01}^{(0)} & K_{02}^{(0)} + K_{01}^{(0)} \\
K_{12}^{(0)} & K_{10}^{(0)} + K_{20}^{(0)} & K_{11}^{(0)} + K_{21}^{(0)} \\
0 & K_{10}^{(1)} & K_{12}^{(1)} + K_{11}^{(1)}
\end{bmatrix}
\begin{bmatrix}
\tilde{\varphi}_0 \\
\tilde{\varphi}_1 \\
\tilde{\varphi}_3 \\
\tilde{\varphi}_4
\end{bmatrix}
= 
\begin{bmatrix}
b_{22}^{(0)} + b_{20}^{(1)} \\
b_{02}^{(0)} + b_{02}^{(1)} \\
b_{12}^{(0)} + b_{12}^{(1)} \\
0
\end{bmatrix}
+ 
\begin{bmatrix}
g_2^{(0)} \\
g_0^{(0)} + g_0^{(1)} \\
g_1^{(0)} + g_1^{(1)}
\end{bmatrix}
$$

(15.204)

Furthermore, we can expand these equations to the full $9 \times 9$ set for all of the nodes, finding that, at this stage in the full assembly,

$$
\begin{bmatrix}
K_{22} & K_{20} & K_{21} \\
K_{02} & K_{00} + K_{01} & K_{02} + K_{01} \\
K_{12} & K_{10} + K_{20} & K_{11} + K_{21} \\
0 & K_{10} & K_{12}
\end{bmatrix}
\begin{bmatrix}
\tilde{\varphi}_0 \\
\tilde{\varphi}_1 \\
\tilde{\varphi}_3 \\
\tilde{\varphi}_4
\end{bmatrix}
= 
\begin{bmatrix}
b_{22} + b_{20} + b_{21} \\
b_{02} + b_{02} + b_{01} \\
b_{12} + b_{12} + b_{11} \\
0
\end{bmatrix}
+ 
\begin{bmatrix}
g_2 \\
g_0 + g_0 \\
g_1 + g_1 \\
0
\end{bmatrix}
$$

(15.205)

\[47\text{We assume that the equations are not contradictory.}\]
The pattern is now clear:

- The element $K_{ij}^{(e)}$ of the $3 \times 3$ matrix $[K^{(e)}]$ contributes additively to the element $K_{lm}$ of the $9 \times 9$ assembled matrix $[K]$, where $l$ is the global number of the node with local number $i$ in element $e$ and $m$ is the global number of the node with local number $j$ in element $e$. For example, $K_{00}^{(0)}$ contributes to $K_{11}$ because local node 0 in element 0 coincides with global node 1; $K_{10}^{(1)}$ contributes to $K_{41}$ because local node 1 of element 1 coincides with global node 4 and local node 0 of element 1 coincides with global node 1.

- The element $b_i^{(e)}$ contributes additively to the element $b_l$, where $l$ is the global number of the node with local number $i$ in element $e$. For example, $b_1^{(0)}$ contributes to $b_3$ because local node 1 in element 0 coincides with global node 3.

- The element $g_i^{(e)}$ contributes additively to the element $g_l$, where $l$ is the global number of the node with local number $i$ in element $e$. For example, $g_2^{(1)}$ contributes to $g_3$ because local node 2 of element 1 coincides with global node 3.

From this point, continuing the assembly to include the contributions from the remaining elements is straightforward but tedious.

### 15.23.5 Incorporating the Boundary Conditions

Apart from the nodal values of $\tilde{\varphi}$, only the $g$’s in Eq. (15.205)—or better in what Eq. (15.205) becomes when all elements have been incorporated—remain unknown. To demonstrate how the boundary conditions either render a priori knowledge of these quantities unnecessary or specify
these quantities explicitly, we begin by noting that the fully assembled vector \( \{ g \} \) is

\[
\{ g \} = \begin{pmatrix}
g_1^{(1)} + g_2^{(2)} + g_3^{(3)} + g_4^{(4)} + g_5^{(5)} + g_6^{(6)} + g_7^{(7)} \\
g_1^{(1)} + g_2^{(1)} + g_2^{(2)} + g_3^{(3)} + g_4^{(4)} + g_5^{(5)} + g_6^{(6)} + g_7^{(7)} \\
g_1^{(1)} + g_2^{(1)} + g_2^{(2)} + g_3^{(3)} + g_4^{(4)} + g_5^{(5)} + g_6^{(6)} + g_7^{(7)} \\
g_1^{(1)} + g_2^{(1)} + g_2^{(2)} + g_3^{(3)} + g_4^{(4)} + g_5^{(5)} + g_6^{(6)} + g_7^{(7)} \\
g_1^{(1)} + g_2^{(1)} + g_2^{(2)} + g_3^{(3)} + g_4^{(4)} + g_5^{(5)} + g_6^{(6)} + g_7^{(7)} \\
\end{pmatrix}
\] (15.206)

Remember, now, that \( g_i^{(e)} \) is given by the line integral

\[
g_i^{(e)} = \oint_{\Gamma^{(e)}} N_i^{(e)} \left( \alpha_x \frac{\partial \hat{\varphi}^{(e)}}{\partial x} \hat{i} + \alpha_y \frac{\partial \hat{\varphi}^{(e)}}{\partial y} \hat{j} \right) \cdot \hat{n}^{(e)} \, dl
\] (15.207)
[see Eq. (15.194)] evaluated counterclockwise around the entirety of element \( e \). Furthermore, notice that the various \( g \)'s that are added together in each row of Eq. (15.206) are the line integrals around each element that has a vertex at the global node with which the row in \( \{ g \} \) is associated.

Let us look at a few representative elements of \( \{ g \} \). Suppose we represent the line integrals by a notation like \((i \to j + j \to k + k \to i)\) for the integral that runs from (global) node \( i \) to (global) node \( j \) to (global) node \( k \) and back to (global) node \( i \) around a particular element. Because the integrand must be continuous along and across the lines separating elements, we need not actually keep track of which element we are traversing on any of the indicated segments. In this notation, for example, \( g_4 \), which is the sum of the line integrals around all the elements having a vertex at global node number 4, becomes

\[
\begin{align*}
(4 \to 3 + 3 \to 1 + 1 \to 4) & \quad + \\
(4 \to 1 + 1 \to 2 + 2 \to 4) & \quad + \\
(4 \to 2 + 2 \to 5 + 5 \to 4) & \quad + \\
(4 \to 5 + 5 \to 7 + 7 \to 4) & \quad + \\
(4 \to 7 + 7 \to 6 + 6 \to 4) & \quad + \\
(4 \to 6 + 6 \to 3 + 3 \to 4) & \quad + 
\end{align*}
\]

Next, note that the middle integral in each line is zero, because the interpolation function along that line is identically zero at all points.\(^{48}\) Then, note that each of the other integrals appears twice, traversed once in each direction, and the two together cancel. This particular sum boils down to zero! The same behavior characterizes all nodes not on the boundary of \( \Omega \). (There is only one such node—global node 4—in the present example.)

We can examine other elements in the vector \( \{ g \} \) using the same shorthand notation. For example, the element \( g_2 \) would be represented by

\[
\begin{align*}
(2 \to 4 + 4 \to 1 + 1 \to 2) & \quad + \\
(2 \to 5 + 5 \to 4 + 4 \to 2) & \quad + 
\end{align*}
\]

As in the previous paragraph, the middle integral in each item is zero and the first integral in the first item cancels the third integral in the second item. Since this node is on the boundary, however,\(^ {48}\)See Exercise 15.15.
complete cancellation does not happen; we are left with $1 \rightarrow 2 + 2 \rightarrow 5$. Similar considerations reduce the vector $\{g\}$ to the much simpler vector

$$
\{g\} = \begin{pmatrix}
3 \rightarrow 0 + 0 \rightarrow 1 \\
0 \rightarrow 1 + 1 \rightarrow 2 \\
1 \rightarrow 2 + 2 \rightarrow 5 \\
6 \rightarrow 3 + 3 \rightarrow 0 \\
0 \rightarrow 1 + 1 \rightarrow 2 \\
2 \rightarrow 5 + 5 \rightarrow 8 \\
7 \rightarrow 6 + 6 \rightarrow 3 \\
8 \rightarrow 7 + 7 \rightarrow 6 \\
5 \rightarrow 8 + 8 \rightarrow 7
\end{pmatrix}
$$

(15.208)

Only integrals along edges that are on the boundary of $\Omega$ remain. Resolution of all remaining unknowns therefore lies in the boundary conditions!

The explicit treatment of these terms from this point depends on what type of boundary conditions are specified. With Dirichlet conditions, the solution is specified on the boundary and the value at each node so constrained is therefore fixed by the boundary conditions. To illustrate the incorporation of this boundary condition, consider a Dirichlet condition at a particular node $m$. We first replace the row corresponding to global node $m$ in the equation $[K]\{\tilde{\varphi}\} = \{b + g\}$ with the equation $\tilde{\varphi}_m = p_m$, where $p_m$ is the value specified for node $m$ by the boundary condition. In short,

- The element in the $m$-th column of the $m$-th row of $[K]$ is replaced by 1,
- All other elements in that row are replaced by 0, and
- The $m$-th element in the vector $\{b + g\}$ is replaced by $p_m$.

49

Then, to complete the incorporation of this boundary condition, we must reflect the influence of this known value $\tilde{\varphi}_m$ on the remaining equations in order to restore the symmetry of the stiffness matrix. As an example, consider the $n$-th equation ($n \neq m$) in the system:

$$
K_{n0} \tilde{\varphi}_0 + K_{n1} \tilde{\varphi}_1 + \cdots + K_{nm} \tilde{\varphi}_m + \cdots + K_{n7} \tilde{\varphi}_7 + K_{n8} \tilde{\varphi}_8 = b_n + g_n
$$

(15.209)

Since $\tilde{\varphi}_m = p_m$, this expression can be recast as

$$
K_{n0} \tilde{\varphi}_0 + K_{n1} \tilde{\varphi}_1 + \cdots + 0 \tilde{\varphi}_m + \cdots + K_{n7} \tilde{\varphi}_7 + K_{n8} \tilde{\varphi}_8 = b_n + g_n - K_{nm} p_m
$$

(15.210)

Such recasting of all equations for $n \neq m$ is accomplished by

- Multiplying by $p_m$ all elements except the element associated with node $m$ in the column associated with node $m$ of the original matrix $[K]$,
- Subtracting each resulting product from the corresponding element in the original vector $\{b + g\}$, and then
- Substituting zero for all elements in the $m$-th column of $[K]$ except the element in the $m$-th row (which had already been set to 1).

So, for each node on which a Dirichlet condition is declared,

- The associated element in the vector $\{b + g\}$ is replaced by the known value at the node corresponding to that element,
- The other elements in the vector $\{b + g\}$ are adjusted as per Eq. (15.210),
- All elements in the associated row and the associated column of $[K]$ except the element in the intersection of that row and column are set to 0, and

49 The original $m$-th equation is then interpreted as an equation giving $g_m$ after the solution to the altered set of equations has been found.
• The element at that intersection is set to 1.

If we have Dirichlet conditions at all bounding nodes, incorporation of these conditions results in replacing the original vector \( \{b + g\} \) with a new vector containing no unknown quantities. The problem is now reduced to solving a fully determined (probably large) set of simultaneous linear equations for the vector \( \{\tilde{\varphi}\} \).

When Neumann conditions are specified, we are given the quantity defined in Eq. (15.72). This gives us enough information to evaluate the integral in Eq. (15.207), not around an entire element but along a portion of the boundary of the region \( \Omega \). We are therefore in a position explicitly to evaluate the integrals symbolized above by \( 3 \rightarrow 0, 0 \rightarrow 1, \) etc. Thus, the \( g \)'s associated with portions of the boundary along which Neumann conditions are specified have values directly determinable from those boundary conditions. If Neumann conditions are specified on all boundaries, all parameters in the equation \( [K]\{\tilde{\varphi}\} = \{b + g\} \) except the dependent variable \( \tilde{\varphi} \) are known from the beginning, so again the problem is reduced to solving a fully determined set of simultaneous linear equations.

We can proceed no further in this example without resorting to a specific coding language, so we postpone further discussion to later sections.

15.26 Using OCTAVE to Solve 2D PDEs via an FEM

15.26.1 A General Coding

The process of assembling the system of equations is tedious, if not difficult. The task of solving the resulting large set of simultaneous algebraic equations is to be sure, straightforward, but it can be enormously time-consuming, especially when the region of interest is more complicated than the simple eight-element geometry that we have so far discussed. The task is clearly one for a computer, and it will almost certainly be carried out numerically, not symbolically. Ideally, we would like to input an equation of the form of Eq. (15.70), a suitable definition of the region of interest, and all applicable boundary conditions, assigning to the computer the tasks of (1) constructing the complete equation begun in Eq. (15.205), (2) applying the boundary conditions, and (3) solving the system of equations.

In order to simplify the computer coding for this example and provide a clearer picture of finite element programming without becoming overwhelmed by geometric details, we will impose several restrictions on the problems that our program can solve. First, we will limit ourselves to a square domain located in the first quadrant with one corner at the origin, such that \( 0 \leq x, y \leq L \). Further, to ensure that all elements will be isosceles right triangles (which simplifies computation considerably), we will divide each edge of this square into the same number \( d \) of segments of equal length \( L/d \). In addition, we will restrict \( \alpha_x, \alpha_y, \beta, \) and \( f \) to be constants and, instead of allowing for general boundary conditions, we will assume the constant Dirichlet conditions

\[
\varphi(x, 0) = p_2 ; \quad \varphi(x, L) = p_1
\]  

(15.211)

on the bottom \((y = 0)\) and top \((y = L)\) boundaries, the linear variation

\[
\varphi(0, y) = \frac{p_1 - p_2}{L} y + p_2
\]  

(15.212)

between the bottom and top values along the left edge \((x = 0)\), and the Neumann boundary condition

\[
\frac{\partial \varphi}{\partial x}(L, y) = q
\]  

(15.213)

with \( q \) constant along the right edge \((x = L)\). Collectively, these assumptions illustrate several different types of boundary condition.
In the first segment of the coding, we will request input of the various parameters describing the problem. Appropriate OCTAVE statements are:

```octave
L = input('Enter length of side (L): ');
d = input('Enter number of segments (d): '); d = fix(d);
alpha_x = input('Enter alpha_x: ');
alpha_y = input('Enter alpha_y: ');
beta = input('Enter beta: ');
f = input('Enter f: ');
p1 = input('Enter value for top edge (p1): ');
p2 = input('Enter value for bottom edge (p2): ');
q = input('Enter q: ');
```

The number of elements, \( M \), is twice the square of the number of segments on each edge, the number of nodes, \( N \), is equal to the square of one plus the number of segments on each edge, i.e., to \((M+1)^2\), and the length of each segment of each edge is the length of the edge divided by the number of segments into which the edge is divided, all as reflected in the additional coding:

```octave
M = 2*d^2; # Calculate number of elements
N = (d+1)^2; # Calculate number of nodes
dx = L/d; # Calculate segment size
```

As illustrated in Fig. 15.22, the number of segments \( d \) also provides a general way to refer to element numbers and global node numbers for varying numbers of elements.

After calculating the number of elements and nodes involved, we find the coordinates of all of the nodes. We elect to store these coordinates in two vectors, \( x \) and \( y \), where each entry is the \( x \) or \( y \) coordinate of the node whose global node number corresponds to the index of the entry in the vector. These assignments can be made with two nested for loops, the inner loop incrementing row numbers down a column of nodes and the outer loop incrementing the column number. In this way the nodes are traversed in the sequence in which we numbered them. Further, we can keep track of the correct global node number with a simple counter. To accomplish the assignment of values to \( x \) and \( y \), we would use the coding:

```octave
x = zeros(1,N); # Create array to store x values
y = zeros(1,N); # Create array to store y values
ct = 0; # Initialize a counter variable
for i = 0:d # Start row number loop
    for j = 0:d # Start column number loop
        ct=ct+1; # Increment counter
        x(ct) = i*dx; # Find x coordinate
        y(ct) = L - j*dx; # Find y coordinate
    end # End loops
end
```

We assign global and local node numbers following the pattern in Fig. 15.23, which differs from Fig. 15.21 only in the incrementation of all numbers by one to reflect the point at which indices start in OCTAVE. The next step is to create and store the connectivity matrix, so we can easily convert between local and global node numbers. As illustrated in Table 15.3, this matrix is a two-dimensional array in which the index of each row is an element number, and the entries in each row are the global node numbers of the first, second, and third local nodes of that element. Then, in developing the coding to create this matrix, we recognize that

50The fix function guarantees that parameters that should be integers are, in fact, integers for the subsequent calculations, regardless of how they happen to be entered. In OCTAVE, all entered values will be floating, so no adjustment is necessary on those parameters.
Figure 15.22: General element and global node numbering based on the number of segments \( d \) into which each edge is divided. Consistent with OCTAVE conventions, we have here numbered elements and nodes starting at 1.

Figure 15.23: Simple square domain based on Fig. 15.20 with element numbers (left and right), global node numbers (left), and local node numbers (right), all starting at 1 as appropriate to the way arrays are indexed in OCTAVE.
Table 15.3: The connectivity matrix when array indices start at 1.

<table>
<thead>
<tr>
<th>Element</th>
<th>Global number of local node 1</th>
<th>Global number of local node 2</th>
<th>Global number of local node 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>9</td>
<td>8</td>
</tr>
</tbody>
</table>

- **Integer** division of the element number minus 1 [i.e., \((e - 1)\)] by \(2 \times d\) yields one less than the index of the vertical line in Fig. 15.22 on which the first node of the element lies. Here, the index of the vertical line is simply a count of the number of vertical lines from the left, so the first vertical line is at index 1, the second at index 2, and the last at index \(d + 1\). (Remember that indices in OCTAVE start at 1.)

- On a given vertical line, the global node number of the first local node increases by **one** every time the element number increases by **two**. Therefore, we can in general find the global node number of an element’s first local node by
  - ○ performing integer division of the element number minus 1 [i.e., \((e - 1)\)] by two,
  - ○ adding one, and
  - ○ adding the index of the vertical line on which the node lies.

- Once the global node number of the first local node of a particular element has been found, the remaining two global node numbers for that element follow easily. The way in which they follow, however, depends on whether the element number is odd or even.

Thus, the connectivity matrix \(cm\) is created by the statements

```matlab
cm = zeros(3,M); % Create 3 by M null array
for e = 1:M % Loop through all elements
    vl = fix((e-1)/(2*d)) + 1; % Find index of vertical line
    cm(1,e) = fix((e-1)/2) + 1 + vl; % Find global number of node 1
    if 2*fix(e/2) == e % If e is even
        cm(2,e) = cm(1,e) + d + 1; % find global number of node 1
        cm(3,e) = cm(1,e) + d; % find global number of node 2
    else % If e is odd
        cm(2,e) = cm(1,e) + d; % find global number of node 1
        cm(3,e) = cm(1,e) - 1; % find global number of node 2
    end
end
```

The next step is to construct the stiffness matrix \([K^{(e)}]\) for a single element. As given in Eq. (15.192), the elements of \([K^{(e)}]\) depend on both the interpolation functions and their derivatives. Since our elements are all isosceles right triangles, the coordinates of all local nodes are related through the leg length, \(\Delta x\). Therefore, we can substitute for \(x_2^{(e)}\) and \(x_3^{(e)}\) in terms of \(x_1^{(e)}\) and \(\Delta x\),

\[\text{51 Several invocations of the } \text{fix command are necessary to make sure certain calculated values are integers.}\]
and for $y_2^{(e)}$ and $y_3^{(e)}$ in terms of $y_1^{(e)}$ and $\Delta x$ in the expression for $K_{ij}^{(e)}$. However, because of the different orientation of odd and even numbered elements, these substitutions are not identical for the two cases. For odd numbered elements, we set

$$x_2^{(e)} = x_1^{(e)} + \Delta x \quad , \quad x_3^{(e)} = x_1^{(e)} \quad \text{and} \quad y_2^{(e)} = y_3^{(e)} = y_1^{(e)} + \Delta x \quad (e \text{ odd}) \quad (15.214)$$

In contrast, for even numbered elements, we set

$$x_2^{(e)} = x_3^{(e)} = x_1^{(e)} + \Delta x \quad , \quad y_2^{(e)} = y_1^{(e)} \quad , \quad y_3^{(e)} = y_1^{(e)} + \Delta x \quad (e \text{ even}) \quad (15.215)$$

With these simplifications, the integrals defining the elements of $[K^{(e)}]$ are much easier to evaluate. Whether the element number $e$ is even or odd, we elect to do the $y$ integral first and the $x$ integral second. In all cases, $x$ will then run from $x_1^{(e)}$ to $x_1^{(e)} + \Delta x$. The limits on $y$, however, will depend on whether $e$ is odd or even. In the present mesh, the triangular elements are all right, isosceles triangles and the diagonal lines defining the hypotenuses of these triangles lie at a 45° angle to the horizontal. Thus, when $e$ is even (see Fig. 15.22), the limits on $y$ will run from $y_1^{(e)}$ to $y_1^{(e)} + x - x_1^{(e)}$, and we conclude that

$$\int \int_{\Omega^{(e)}} \ldots \ dx \ dy = \int_{x_1^{(e)}}^{x_1^{(e)}+\Delta x} \left[ \int_{y_1^{(e)}}^{y_1^{(e)}+x-x_1^{(e)}} \ldots \ dy \right] dx \quad (e \text{ even}) \quad (15.216)$$

When $e$ is odd, on the other hand, the limits on $y$ will instead run from $y = y_1^{(e)} + x - x_1^{(e)}$ to $y_1^{(e)} + \Delta x$, and we conclude that

$$\int \int_{\Omega^{(e)}} \ldots \ dx \ dy = \int_{x_1^{(e)}}^{x_1^{(e)}+\Delta x} \left[ \int_{y_1^{(e)}}^{y_1^{(e)}+x-x_1^{(e)}} \ldots \ dy \right] dx \quad (e \text{ odd}) \quad (15.217)$$

Because the integrands are no worse than quadratic in $x$ and $y$, their evaluation is straightforward, but there are many of them and we elect to invoke a symbol manipulating program.\(^{52}\) It turns out that the entries in $[K^{(e)}]$ depend only on the parameters $\Delta x$, $\alpha_x$, and $\alpha_y$, and not upon $e$, the specific element. In other words, there exist only two distinct elemental stiffness matrices, one for even elements and one for odd elements, namely

$$K_{ij}^{(e)}_{\text{even}} = \begin{bmatrix}
\frac{\beta \Delta x^2}{12} + 6\alpha_x & \frac{\beta \Delta x^2}{24} - 12\alpha_x & \frac{\beta \Delta x^2}{24} \\
\frac{\beta \Delta x^2}{24} - 12\alpha_x & \frac{\beta \Delta x^2}{24} + 6\alpha_x + 6\alpha_y & \frac{\beta \Delta x^2}{24} - 12\alpha_y \\
\frac{\beta \Delta x^2}{24} & \frac{\beta \Delta x^2}{24} - 12\alpha_y & \frac{\beta \Delta x^2}{12} + 6\alpha_x + 6\alpha_y \\
\end{bmatrix}$$

and

$$K_{ij}^{(e)}_{\text{odd}} = \begin{bmatrix}
\frac{\beta \Delta x^2}{12} + 6\alpha_y & \frac{\beta \Delta x^2}{24} & \frac{\beta \Delta x^2}{24} - 12\alpha_y \\
\frac{\beta \Delta x^2}{24} & \frac{\beta \Delta x^2}{24} + 6\alpha_x & \frac{\beta \Delta x^2}{24} - 12\alpha_x \\
\frac{\beta \Delta x^2}{24} - 12\alpha_y & \frac{\beta \Delta x^2}{24} - 12\alpha_x & \frac{\beta \Delta x^2}{12} + 6\alpha_x + 6\alpha_y \\
\end{bmatrix}$$

The coding that constructs these two matrices in our program is

\(^{52}\)Details of this evaluation are described more fully in Appendix 15.F.
Kodd = zeros(3,3);  # Create two 3 by 3 arrays
Keven = zeros(3,3);

bx = beta*dx^2;  # Evaluate a common quantity

Keven(1,1) = (bx + 6*alpha_x)/12;  # Assign the appropriate value to
Keven(2,2) = (bx + 6*alpha_x + 6*alpha_y)/12;  # each K(i,j). Note that the array
Keven(3,3) = (bx + 6*alpha_y)/12;  # are symmetric, and that Kodd
Keven(1,2) = (bx - 12*alpha_x)/24;  # includes all of the same values as
Keven(2,1) = Keven(1,2);  # Keven, but in different locations.
Keven(2,3) = (bx - 12*alpha_y)/24;
Keven(3,2) = Keven(2,3);
Keven(1,3) = bx/24;
Keven(3,1) = bx/24;

Kodd(1,1) = Keven(3,3);
Kodd(2,2) = Keven(1,1);
Kodd(3,3) = Keven(2,2);
Kodd(1,2) = Keven(1,3);
Kodd(2,1) = Kodd(1,2);
Kodd(2,3) = Keven(1,2);
Kodd(3,2) = Kodd(2,3);
Kodd(1,3) = Keven(2,3);
Kodd(3,1) = Kodd(1,3);

We are now ready to assemble the complete matrix $[K]$, some of whose elements were shown in Eq. (15.205). At the same time, we can also construct the vector $\{b\}$, whose entries $b_i(e)$ can be found by using a symbol manipulating program to evaluate the integrals in Eq. (15.193). It turns out that all three elements $b_1(e)$, $b_2(e)$, and $b_3(e)$ have the same value $f \Delta x^2/6$, regardless of whether $e$ is odd or even. Following the algorithm described at the end of Section 15.23.4 for assembling the final stiffness matrix, we start by creating a null matrix $[K]$ and a null vector $\{b\}$ of the proper dimensions. Then, we step through the elements one at a time, at each step adding the contributions of the element to the accumulating entries at the proper positions in $[K]$ and $\{b\}$. This end is achieved with the coding

```plaintext
K = zeros(N,N);  # Create arrays to store values of
b = zeros(N,1);  # K(i,j) and b(i)

for e = 1:M  # Count through element numbers
  for i = 1:3  # For each local node of an element
    b(cm(i,e)) = b(cm(i,e)) + f*dx^2/6;  # place its contributions at the
    for j = 1:3  # correct locations in [K] and {b}
      if 2*fix(e/2) ~= e
        K(cm(i,e),cm(j,e)) = K(cm(i,e),cm(j,e)) + Kodd(i,j);
      else
        K(cm(i,e),cm(j,e)) = K(cm(i,e),cm(j,e)) + Keven(i,j);
      end
    end
  end
end

53See again Appendix 15.F.
```
Application of the boundary conditions will complete the setup of the problem. To apply the Dirichlet conditions to the top, bottom, and left edges we must first find an algorithm to identify which nodes are on these edges for various values of \( d \). By inspection, we find that, by letting \( i \) run from zero to \( d \), global node numbers on the left boundary are given by \( i+1 \), on the top by \( i(d+1) + 1 \), and on the bottom by \( (i+1)(d+1) \). Once we know these numbers, we can replace the appropriate equations by the given values of \( \tilde{\varphi} \), and then reflect the influence of this relationship in the other equations as described in Section 15.23.5. The requisite coding is

```matlab
for i = 0:d
    u = i + 1; # Nodes on the left boundary
    s = i*(d+1) + 1; # Nodes on the top boundary
    t = (i+1)*(d+1); # Nodes on the bottom boundary
    p3 = (p1-p2)/L * y(u) + p2; # Find values of phi on left boundary
    for j = 1:N
        K(j,s) = 0; # Set rows in K to zero where value # of phi is known
        K(j,t) = 0; # Set values of b
        K(j,u) = 0;
        b(s) = p1; # Set values of b
        b(t) = p2;
        b(u) = p3;
        if j ~= s # Reflect influence
            b(j) = b(j) - K(s,j)*p1; # of known values of
        end # phi in the other
        K(s,j) = 0; # equations
        if j ~= t
            b(j) = b(j) - K(t,j)*p2;
        end
        K(t,j) = 0;
        if j ~= u
            b(j) = b(j) - K(u,j)*p3;
        end
        K(u,j) = 0;
    end
    K(s,s) = 1; % Set the appropriate entry to 1 in # the rows where phi is known
    K(t,t) = 1;
    K(u,u) = 1;
end
```

Since the Neumann condition is specified only on the right edge, the unit normal vector \( \hat{n} \) in Eqs. (15.72) and (15.194) becomes simply \( \hat{i} \). Also, as shown in Section 15.23.5, we need worry only about segments on the boundary of \( \Omega \) when integrating to find \( g^\prime_i \). These two items and the fact that all elements with boundaries on the right edge have even numbers reduce Eq. (15.194) to

\[
 g^{(e)}_i = \int_{y_i^{(e)}}^{y_i^{(e)}+\Delta x} N^{(e)}_i \alpha x \frac{\partial \tilde{\varphi}}{\partial x} dy
\]  

(15.220)

Since we know that \( \alpha x (\partial \tilde{\varphi}/\partial x) = q \) and that \( x = x_0^{(e)} + \Delta x \), we can easily evaluate this expression, provided we assume \( \alpha x \) can be treated as a constant over the area of each element. Again invoking a symbol manipulating program,\(^{54}\) we find that

\[
 g_1^{(e)} = 0 ; \quad g_2^{(e)} = \frac{q \Delta x}{2} ; \quad g_3^{(e)} = \frac{q \Delta x}{2}
\]  

(15.221)

\(^{54}\)See Appendix 15.F once more.
The contribution of each node to the complete \( \{g\} \) vector will be twice this value, since it consists of two parts, one from each right boundary element that contains the node. We don’t need \( g^{(e)}_1 \) since no node on the right edge is a first local node. The necessary additional coding, including the consolidation of the inhomogeneities into a single vector, is

```octave
% Create vector to store g values
for i = 1:d-1
    nd = d*(d+1)+i+1; % Nodes on right boundary
    g(nd) = q * dx;
end
b = b+g; % Store the vector \{b+g\} in \{b\}
```

At this point, \( K \) contains the coefficient matrix and \( b \) the inhomogeneities for the system of simultaneous linear equations we wish to solve.

We need only solve the system and then write the results into an appropriate two-dimensional array. The remaining coding thus is

```octave
phi = K\b; % Solve the equation K*phi = b
# writing the results to phi
A = zeros(d+1,d+1); % Create a d+1 by d+1 array
cnt=0; % Initialize a counter
for i = 1:d+1 % Use nested for loops to write all entries in A
    for j = d+1:-1:1
        cnt=cnt+1; % Increment counter
        A(i,j) = phi(cnt);
    end % End inner loop
end % End outer loop
```

Since the nodes are, in this example, uniformly spaced over the region of the problem, we need not be concerned about knowing the nodal coordinates for purposes of plotting the solution. The coordinates of the node corresponding to each value in the two-dimensional array \( A \) are proportional to the indices of that value in the array.

All of the preceding OCTAVE code has been incorporated into the procedure fem2d.m, a listing of which can be found in Appendix 15.G.3. (The file itself can be copied from the directory $HEAD/octave.) When run, the procedure simply asks for necessary input, calculates the solution, and returns that solution in the array \( A \). The array \( A \) can then be examined, probably by plotting a surface, to view the solution.

### 15.26.2 An Example: Isotropic Heat Flow

We will now demonstrate fem2d by applying it to an equation in isotropic heat flow. The steady-state temperature \( u(x, y) \) in a square plate of uniform composition is a solution to the two-dimensional Laplace equation

\[
\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0
\]

(15.222)

to which Eq. (15.70) reduces when we replace \( \tilde{\varphi} \) with \( u \) and set

\[
\alpha_x = 1 \quad ; \quad \alpha_y = 1 \quad ; \quad \beta = 0 \quad ; \quad f = 0
\]

(15.223)
We will take the plate to be ten units square \((L = 10)\), and position it in the first quadrant with a corner at the origin, and we will apply the boundary conditions

\[
\begin{align*}
  u(x, 0) &= 0 ; \quad u(x, 10) = 100 ; \quad u(0, y) = 10y ; \quad \frac{\partial u}{\partial x}(10, y) = 0
\end{align*}
\]

i.e., \(p_2 = 0, p_1 = 100, L = 10\), and \(q = 0\). With these choices, we run the program and solve the problem with the single statement

\[
fem2d
\]

At the several prompts generated by this statement, we enter the values

\[
\begin{align*}
  \text{Enter length of side (L):} & \quad 10 \\
  \text{Enter number of segments (d):} & \quad 8 \\
  \text{Enter alpha}_x & \quad 1 \\
  \text{Enter alpha}_y & \quad 1 \\
  \text{Enter beta} & \quad 0 \\
  \text{Enter f} & \quad 0 \\
  \text{Enter value for top edge} & \quad 100 \\
  \text{Enter value for bottom edge} & \quad 0 \\
  \text{Enter q} & \quad 0
\end{align*}
\]

Each line appears one at a time, and waits for us to input the requested parameter. After the last parameter has been entered, the program generates a solution and stores it in the array \(A\), which is returned to the main program and can easily be displayed by invoking the statements

\[
\begin{align*}
  [x, y] &= \text{meshgrid}(0:1.25:10, 0:1.25:10); \\
  \text{mesh}(x, y, A', \text{'edgecolor'}, \text{'black'}, \text{'linewidth'}, 3); \\
  \text{set(gca,'fontsize',12)} \\
  \text{xlabel('x','fontsize',20)} \\
  \text{ylabel('y','fontsize',20)} \\
  \text{view([45.0,20.0])}
\end{align*}
\]

Here, transposing \(A\) and adjusting the view orients the figure properly with respect to the \(x\) and \(y\) axes and orients the axes on the page so that it is easy to associate the graph with the solution as we have displayed it as a matrix. The resulting graph is shown in Fig. 15.24. This figure shows that the temperature varies linearly throughout the plate. This result is, of course, entirely consistent with our intuition: the temperature in the plate should vary linearly between the extremes on opposite edges. The printed solution obtained with the statement\(^{55}\)

\[
\begin{align*}
  \text{rot90(} A \text{)} \\
  \text{ans =} & \quad \begin{bmatrix}
    100.00 & 100.00 & 100.00 & 100.00 & 100.00 & 100.00 & 100.00 & 100.00 & 100.00 \\
    87.50 & 87.50 & 87.50 & 87.50 & 87.50 & 87.50 & 87.50 & 87.50 & 87.50 \\
    75.00 & 75.00 & 75.00 & 75.00 & 75.00 & 75.00 & 75.00 & 75.00 & 75.00 \\
    62.50 & 62.50 & 62.50 & 62.50 & 62.50 & 62.50 & 62.50 & 62.50 & 62.50 \\
    50.00 & 50.00 & 50.00 & 50.00 & 50.00 & 50.00 & 50.00 & 50.00 & 50.00 \\
    37.50 & 37.50 & 37.50 & 37.50 & 37.50 & 37.50 & 37.50 & 37.50 & 37.50 \\
    25.00 & 25.00 & 25.00 & 25.00 & 25.00 & 25.00 & 25.00 & 25.00 & 25.00 \\
    12.50 & 12.50 & 12.50 & 12.50 & 12.50 & 12.50 & 12.50 & 12.50 & 12.50 \\
    0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00
  \end{bmatrix}
\end{align*}
\]

reveals even more clearly this anticipated linearity.

\(^{55}\)Here, we rotate the matrix 90° counterclockwise to orient it the way we have visualized the region in which a solution is sought. We have also deleted three zeroes from the end of every value so that the matrix will fit more comfortably on the page.
When working with numerical approximation techniques such as finite element analysis, we must always be wary of the accuracy of results. When investigating a problem with a less evident solution, we would be tempted to test the results by performing additional analyses using a larger number of elements. In the present case, for example, we could generate the solution on a finer grid by specifying the values

```python
fem2d
Enter length of side (L): 10
Enter number of segments (d): 16
Enter alpha_x: 1
Enter alpha_y: 1
Enter beta: 0
Enter f: 0
Enter value for top edge: 100
Enter value for bottom edge: 0
Enter q: 0
```

which will generate a solution on a $17 \times 17$ grid with twice as many divisions as the first solution. Then, we might extract from that more refined solution the solution on the $9 \times 9$ grid with the statements

```python
B = A(1:2:17, 1:2:17);
rot90(B)
```

```
ans =
    100.00    100.00    100.00    100.00    100.00    100.00    100.00    100.00    100.00
    87.50    87.50    87.50    87.50    87.50    87.50    87.50    87.50    87.50
    75.00    75.00    75.00    75.00    75.00    75.00    75.00    75.00    75.00
    62.50    62.50    62.50    62.50    62.50    62.50    62.50    62.50    62.50
    50.00    50.00    50.00    50.00    50.00    50.00    50.00    50.00    50.00
    37.50    37.50    37.50    37.50    37.50    37.50    37.50    37.50    37.50
```
We discover that, at least to two digits after the decimal place, the two solutions are identical, and we conclude that discretization and roundoff error are under control to this level of accuracy. (Because we have solved the algebraic equations by a direct method, we need not here be concerned about convergence error.)

15.30 Exercises

Note: In these exercises, we refer to the programs developed in the text without appending the file type, adopting this approach to leave to you the choice of which language to use in addressing the exercise.

15.1. Recast \texttt{fdm1d} to solve the ODE of Eq. (pde:diffeq) when \( f(x) = kx \), the boundary conditions are
\[
\frac{\partial \varphi}{\partial x}(0) = p ; \quad \varphi(L) = q
\]
and \( \alpha, \beta, k, p, q, \) and \( L \) are constants whose values are to be read in at execution time. Then, using your command file, explore the solution to the equation for various values of \( k \) when
\[
\alpha = -4.0 ; \quad \beta = 4.0 ; \quad L = 10.0 ; \quad p = q = 0
\]
Optional: (a) Find an analytic solution to the equation in this exercise and compare the exact results with the approximate results generated by your modification of \texttt{fdm1d}. (b) Find the points at which the solution has the value zero, both starting with the solution obtained in the main exercise and working from the exact solution obtained in optional part (a).

15.2. Suppose the one-dimensional string of length \( l \) discussed in Section 15.1.1 hangs vertically and is acted on by gravity. Suppose that \( u(x, t) \) and \( v(x, t) \) give the transverse (horizontal) and longitudinal (vertical) displacements of the particle of the string nominally located at \( x \), which is measured downward from the top of the string. (a) Examine the forces acting on this string, deduce the general equations for both longitudinal and transverse motion of the string, and show ultimately that, if the motion is entirely transverse and the amplitude of the motion is small, the equations reduce to
\[
\rho(x) \frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x} \left( \tau(x) \frac{\partial u}{\partial x} \right) ; \quad \tau(x) = -g \int_x^l \rho(x') dx'
\]
where \( \rho(x) \) is the mass per unit length of the string. (b) Taking \( \rho \) to be constant, show that the tension \( \tau(x) \), which is simply equal to the weight of the string below \( x \), is given by \( \tau(x) = pg(l-x) \). With this restriction, the equation of motion then becomes
\[
\frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x} \left( (l-x) g \frac{\partial u}{\partial x} \right)
\]
(c) Recast this equation in dimensionless form by introducing the variables \( \overline{x} = x/l \) and \( \overline{t} = t\sqrt{g/l} \). (d) Suppose you seek a sinusoidal solution for which \( u(\overline{x}, \overline{t}) = f(\overline{x}) \cos \omega \overline{t} \). Find the ODE satisfied by \( f(\overline{x}) \) and then introduce the new independent variable \( y \) defined by \( y^2 = 1 - \overline{x} \) to find that, expressed as a function of \( y, f \) satisfies the Bessel equation
\[
y^2 \frac{d^2 f}{dy^2} + y \frac{df}{dy} - 4\omega^2 y^2 f = 0
\]
Note: The variable transformation \( y^2 = 1 - x \) in effect recognizes that the vertical string is more appropriately treated by locating points on the string relative to the bottom rather than the top of the string!
15.3. Show that every second-order linear ODE can be cast in self-adjoint form, i.e., show that functions \(a(x), \gamma(x),\) and \(f(x)\) can be found such that the general second-order linear ODE
\[
a(x) \frac{d^2 \varphi}{dx^2} + b(x) \frac{d \varphi}{dx} + c(x) \varphi = g(x)
\]
can be recast in the form of Eq. (15.67) in Section 15.1.7. Hint: Multiply the equation by an undetermined integrating factor \(h(x),\) find a (first-order) differential equation for \(h(x),\) and solve that equation—at least formally.

15.4. Recast \(\text{fimd}\) to solve the equation
\[
\frac{d^2 u}{dx^2} + kxu = 0
\]
over the interval \(0 \leq x \leq L\) subject to the boundary conditions \(u(0) = 0.0\) and \(du(L)/dx = 1.0,\) arranging your program so the value of \(k\) is entered at execution time. Then, explore the solution in some detail for various values of \(k,\) making sure to assess the accuracy of your solution.

15.5. In (two-dimensional) polar coordinates \((\nabla, \phi),\) the Laplace equation is
\[
\nabla^2 u = \frac{1}{\rho} \frac{\partial}{\partial \rho} \left( \nabla u \right) + \frac{1}{\rho^2} \frac{\partial^2 u}{\partial \varphi^2} = 0 \quad \Rightarrow \quad \nabla^2 \frac{\partial u}{\partial \rho} + \nabla \frac{\partial u}{\partial \varphi} + \frac{\partial^2 u}{\partial \rho^2} = 0
\]
Suppose a solution is sought in the region \(0 < a \leq \rho \leq b,\) \(0 \leq \phi \leq 2\pi\) subject to the boundary conditions
\[
u(a, \phi) = f(\phi) \quad ; \quad \nu(b, \phi) = g(\phi)
\]
and the requirement that the solution be periodic with period \(2\pi\) in \(\phi.\) Let \(\Delta \nabla = (b - a)/n\) and \(\Delta \phi = 2\pi/m\) and then let \(\nabla_i = a + i \Delta \nabla, i = 0, 1, 2, \ldots, n\) and \(\phi_j = j \Delta \phi, j = 0, 1, 2, \ldots, m.\) Finally let \(u_{i,j} = u(\nabla_i, \phi_j).\) Discretize this equation. In particular use a central difference formula to approximate the derivative \(\partial \nu/\partial \varphi.\) You should find ultimately that
\[
u_{i,j} = \frac{u_{i+1,j} \left( \nabla_i \Delta \phi^2 + \frac{1}{2} \nabla_i \Delta \phi^2 \Delta \nabla \right) + u_{i-1,j} \left( \nabla_i \Delta \phi^2 - \frac{1}{2} \nabla_i \Delta \phi^2 \Delta \nabla \right) + \left( u_{i,j+1} + u_{i,j-1} \right) \Delta \nabla^2}{2 \nabla_i^2 \Delta \phi^2 + 2 \Delta \nabla^2}
\]
Determine how this equation must be modified when \(i = 0, i = n\) and/or \(j = 0, j = m,\) and then write a program to solve this equation when
\[
f(\phi) = 0 \quad ; \quad g(\phi) = \begin{cases} & -100 \quad \text{if } -\pi < \phi < 0 \\ & 100 \quad \text{if } 0 \leq \phi \leq \pi \end{cases}
\]
Finally, compile, link, and run your program to explore the solution in some detail.

15.6. Suppose you know three points \((x_0, f_0), (x_1 = x_0 + \Delta x, f_1),\) and \((x_2 = x_0 + 2 \Delta x, f_2)\) on a curve and you wish to estimate the derivative of the corresponding function at \(x = x_0\) by a method that is more accurate than simply the forward difference formula \((f_1 - f_0)/\Delta x.\) You might fit the parabola \(f(x) = Ax^2 + Bx + C\) through the three points and then approximate the derivative of the actual function as the derivative of this parabola at the point \(x = x_0.\) Show that this approach yields the formula
\[
\frac{df}{dx} \bigg|_{x = x_0} = \frac{-3f_0 + 4f_1 - f_2}{2\Delta x}
\]
which is a higher-order forward difference approximation than the one used in the text. Hint: Set up the three equations \(f_1 = Ax_1^2 + Bx_1 + C,\) solve those equations for \(A, B,\) and \(C,\) and note that \(df/dx = 2Ax + B.\) You may find a symbolic manipulating program of substantial assistance.

15.7. Consider the differential equation and boundary conditions
\[
\frac{d^2 u}{dx^2} + k^2 u = 0 \quad ; \quad u(0) = u(L) = 0
\]
(a) Find the analytic solution to this problem and identify the special values of \(k\) that permit non-trivial solutions. (b) Show that the finite difference approach to the problem leads to the matrix
eigenvalue problem \( Au = -k^2 L^2 u / N^2 \) where \( N \) is the number of equal-length segments into which the length \( L \) of the domain is divided and

\[
A = \begin{pmatrix}
-2 & 1 & 0 & \cdots & 0 & 0 & 0 \\
1 & -2 & 1 & \cdots & 0 & 0 & 0 \\
0 & 1 & -2 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & -2 & 1 & 0 \\
0 & 0 & 0 & \cdots & 1 & -2 & 1 \\
0 & 0 & 0 & \cdots & 0 & 1 & -2 \\
\end{pmatrix}
\]

\[
u = \begin{pmatrix}
u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{N-3} \\ u_{N-2} \\ u_{N-1} \\ \end{pmatrix}
\]

Note that, when the domain \( 0 \leq x \leq L \) is divided into \( N \) segments, there will be \( N + 1 \) nodes ranging from \( x_0 = 0 \) to \( x_{N+1} = L \). Because, along the way to a solution, the boundary conditions result in the rows and columns associated with these two nodes being deleted, these matrices will have only \( N - 1 \) rows and columns. (c) Taking \( N = 100 \), use an available numeric processing program like IDL, MATLAB, OCTAVE, or PYTHON to find the first several eigenvalues \( k_n \) and compare your results with the values found in part (a).

15.8. Consider the differential equation and boundary conditions

\[
d^2u/dx^2 + k^2 u = 0 \quad ; \quad u(0) = u(L) = 0
\]

(a) Find the analytic solution to this problem and identify the special values of \( k \) that permit non-trivial solutions. (b) Show that the finite element approach to the problem leads to the generalized eigenvalue problem

\[
Au = -k^2 L^2 Bu / 6N^2
\]

where \( N \) is the number of equal-length elements into which the length \( L \) of the domain is divided and

\[
A = \begin{pmatrix}
-2 & 1 & 0 & \cdots & 0 & 0 & 0 \\
1 & -2 & 1 & \cdots & 0 & 0 & 0 \\
0 & 1 & -2 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & -2 & 1 & 0 \\
0 & 0 & 0 & \cdots & 1 & -2 & 1 \\
0 & 0 & 0 & \cdots & 0 & 1 & -2 \\
\end{pmatrix} \quad ; \quad B = \begin{pmatrix}
4 & 1 & 0 & \cdots & 0 & 0 & 0 \\
1 & 4 & 1 & \cdots & 0 & 0 & 0 \\
0 & 1 & 4 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 4 & 1 & 0 \\
0 & 0 & 0 & \cdots & 1 & 4 & 1 \\
0 & 0 & 0 & \cdots & 0 & 1 & 4 \\
\end{pmatrix}
\]

and

\[
u = \begin{pmatrix}
u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{N-3} \\ u_{N-2} \\ u_{N-1} \\ \end{pmatrix}
\]

Note that, when the domain \( 0 \leq x \leq L \) is divided into \( N \) segments, there will be \( N + 1 \) nodes ranging from \( x_0 = 0 \) to \( x_{N+1} = L \). Because, along the way to a solution, the boundary conditions result in the rows and columns associated with these two nodes being deleted, these matrices will have only \( N - 1 \) rows and columns. (c) Taking \( N = 100 \), use an available numeric processing program like IDL, MATLAB, OCTAVE, or PYTHON to find the first several eigenvalues \( k_n \) and compare your results with the values found in part (a).

15.9. When \( \alpha, \beta, \) and \( f \) are constants, Eqs. (15.67), (15.68), and (15.69) can be solved analytically. Show, for example, that the solution to this boundary value problem is given by

\[
\varphi(x) = A \sin \lambda x + \left( p - \frac{f}{\beta} \right) \cos \lambda x + \frac{f}{\beta}
\]

where

\[
A = \frac{(q - \gamma f / \beta) + (p - f / \beta)(\alpha \lambda \sin \lambda L - \gamma \cos \lambda L)}{\alpha \lambda \cos \lambda L + \gamma \sin \lambda L}
\]
and \( \lambda = \sqrt{-\beta/\alpha} \) when \( \beta/\alpha < 0 \). Using graphical displays in particular, compare the analytic solution in this exercise with the solution obtained numerically by finite difference and finite element approaches. Optional: Find corresponding solutions when \( \beta/\alpha = 0 \) and \( \beta/\alpha > 0 \).

**15.10.** Suppose a linear element \( e \) is characterized by three nodes at \( x^{(e)}_1, x^{(e)}_2, \) and \( x^{(e)}_3 \). Further, let the solution \( \tilde{\varphi}^{(e)} \) on that element be approximated by the quadratic function

\[
\tilde{\varphi}^{(e)} = a^{(e)} + b^{(e)} x + c^{(e)} x^2
\]

(a) Find the constants \( a^{(e)}, b^{(e)}, \) and \( c^{(e)} \) that will make this function match the specific values \( \tilde{\varphi}_1^{(e)}, \tilde{\varphi}_2^{(e)}, \) and \( \tilde{\varphi}_3^{(e)} \) at the points \( x = x^{(e)}_1, x^{(e)}_2, \) and \( x^{(e)}_3 \), respectively. (b) Substituting these values into the above expression and grouping terms appropriately, cast the result in the form

\[
\tilde{\varphi}^{(e)} = \sum_{i=1}^{3} N_i^{(e)}(x)
\]

and show that the shape functions \( N_i^{(e)}(x) \) appropriate to this three-noded linear element are

\[
\begin{align*}
N_1^{(e)}(x) &= \frac{1}{\Delta} \left| \begin{array}{ccc} 1 & x & x^2 \\
x_1^{(e)} & (x_1^{(e)})^2 & (x_1^{(e)})^3 \\
x_2^{(e)} & (x_2^{(e)})^2 & (x_2^{(e)})^3 \\
x_3^{(e)} & (x_3^{(e)})^2 & (x_3^{(e)})^3 \\
\end{array} \right| = \frac{(x_2^{(e)} - x)(x_3^{(e)} - x)}{(x_2^{(e)} - x_1^{(e)})(x_3^{(e)} - x_1^{(e)})} \\
N_2^{(e)}(x) &= \frac{1}{\Delta} \left| \begin{array}{ccc} 1 & x & x^2 \\
x_1^{(e)} & (x_1^{(e)})^2 & (x_1^{(e)})^3 \\
x_2^{(e)} & (x_2^{(e)})^2 & (x_2^{(e)})^3 \\
x_3^{(e)} & (x_3^{(e)})^2 & (x_3^{(e)})^3 \\
\end{array} \right| = \frac{(x - x_1^{(e)})(x_3^{(e)} - x)}{(x_2^{(e)} - x_1^{(e)})(x_3^{(e)} - x_2^{(e)})} \\
N_3^{(e)}(x) &= \frac{1}{\Delta} \left| \begin{array}{ccc} 1 & x & x^2 \\
x_1^{(e)} & (x_1^{(e)})^2 & (x_1^{(e)})^3 \\
x_2^{(e)} & (x_2^{(e)})^2 & (x_2^{(e)})^3 \\
x_3^{(e)} & (x_3^{(e)})^2 & (x_3^{(e)})^3 \\
\end{array} \right| = \frac{(x - x_1^{(e)})(x_2^{(e)} - x)}{(x_3^{(e)} - x_1^{(e)})(x_2^{(e)} - x_2^{(e)})}
\end{align*}
\]

where

\[
\Delta = \left| \begin{array}{ccc} 1 & x_1^{(e)} & (x_1^{(e)})^2 \\
x_1^{(e)} & (x_1^{(e)})^2 & (x_1^{(e)})^3 \\
x_2^{(e)} & (x_2^{(e)})^2 & (x_2^{(e)})^3 \\
x_3^{(e)} & (x_3^{(e)})^2 & (x_3^{(e)})^3 \\
\end{array} \right|
\]

Finally, (c) show that—with \( \xi = (x - x_1^{(e)})/l^{(e)} \)—these functions reduce to

\[
N_1^{(e)}(x) = 2(\xi - 1) \left( \xi - \frac{1}{2} \right) ; \quad N_2^{(e)}(x) = 4\xi(1 - \xi) ; \quad N_3^{(e)}(x) = 2\xi \left( \xi - \frac{1}{2} \right)
\]

when \( x_2^{(e)} \) is midway between \( x_1^{(e)} \) and \( x_3^{(e)} \), i.e., when \( x_2^{(e)} = x_1^{(e)} + \frac{1}{2} l^{(e)} \) and \( x_3^{(e)} = x_1^{(e)} + l^{(e)} \), and

(d) obtain graphs of these three shape functions over the interval \( 0 < \xi < 1 \). Hint: You may find a symbolic manipulating program to be useful at many points in this problem.

**15.11.** Accepting the shape functions given in part (c) of Exercise 15.10, (a) find the \( 3 \times 3 \) matrices

\[
^1K_{ij}^{(e)} = \int_{x_1^{(e)}}^{x_3^{(e)}} dN_i^{(e)} \frac{dN_j^{(e)}}{dx} dx \quad \text{and} \quad ^2K_{ij}^{(e)} = \int_{x_1^{(e)}}^{x_3^{(e)}} N_i^{(e)} N_j^{(e)} dx
\]

and then construct the \( 3 \times 3 \) matrix whose elements are

\[
K_{ij}^{(e)} = a^{(e)} \cdot ^1K_{ij}^{(e)} + b^{(e)} \cdot ^2K_{ij}^{(e)}
\]

to produce the (elemental) stiffness matrix analogous to Eq. (15.98). (b) Find the three-element vector whose elements are

\[
b_i^{(e)} = f^{(e)} \int_{x_1^{(e)}}^{x_3^{(e)}} N_i^{(e)} dx
\]

analogous to Eq. (15.99) and the three element vector whose elements are

\[
g_i^{(e)} = \alpha^{(e)} N_i^{(e)} \left. \frac{d^2}{dx^2} \right|_{x=x_3^{(e)}} - \alpha^{(e)} N_i^{(e)} \left. \frac{d}{dx} \right|_{x=x_1^{(e)}}.
\]
15.30. EXERCISES

Consider a four-node element in two dimensions. Let the nodes be at $(x_0, y_0), (x_1, y_1), (x_2, y_2)$, and $(x_3, y_3)$, and take the interpolating function to be $\tilde{\varphi} = ax + by + cy + dxy$. (a) Find the four shape

\[ \begin{bmatrix} K_{11}^{(e)} & K_{12}^{(e)} & K_{13}^{(e)} \\ K_{21}^{(e)} & K_{22}^{(e)} & K_{23}^{(e)} \\ K_{31}^{(e)} & K_{32}^{(e)} & K_{33}^{(e)} \end{bmatrix} \begin{bmatrix} \tilde{\varphi}_1^{(e)} \\ \tilde{\varphi}_2^{(e)} \\ \tilde{\varphi}_3^{(e)} \end{bmatrix} = \begin{bmatrix} b_1^{(e)} \\ b_2^{(e)} \\ b_3^{(e)} \end{bmatrix} + \begin{bmatrix} g_1^{(e)} \\ 0 \\ g_3^{(e)} \end{bmatrix} \]

where

\[ g_1^{(e)} = -\alpha^{(e)} \frac{d\tilde{\varphi}_1^{(e)}}{dx} \bigg|_{x_1^{(e)}} ; \quad g_3^{(e)} = \alpha^{(e)} \frac{d\tilde{\varphi}_3^{(e)}}{dx} \bigg|_{x_3^{(e)}} . \]

Except for factors of $\alpha^{(e)}, \beta^{(e)}, f^{(e)},$ and $l^{(e)}$, you should have explicit values for $K_{ij}^{(e)}$ and $b_i^{(e)}$.

Hints: (1) Begin by recasting the integrals as integrals on the variable $\xi$. (2) You may find a symbolic manipulating program to be useful at many points in this problem. (3) Assume as in Exercise 15.10 that the point $x = x_2$ is midway between $x_1$ and $x_3$.

15.12. Continuing with the circumstances of the previous two problems, suppose the region of interest is divided into three three-noded elements, $e = 1, 2, 3$ with global nodes 1, 2, 3, 4, 5, 6, 7, nodes 1, 2, and 3 in element 1, nodes 3, 4, and 5 in element 2, and nodes 5, 6, and 7 in element 3. Suppose node 2 is midway between nodes 1 and 3, node 4 is midway between nodes 3 and 5, and node 6 is midway between nodes 5 and 7, but do not suppose the lengths of the three elements are the same. Following the pattern in Section 15.9.4, assemble the elemental equations for these three elements into an equation for the whole system analogous to Eq. (15.130) if the solution is required to satisfy the boundary conditions of Eqs. (15.68) and (15.69).

15.13. Recast fem1d so it will generate a solution to Eq. (15.67) when

- $\alpha$ is a (positive) constant and $\beta = 0$,
- $f(x)$ varies with position in accordance with $f(x) = Ae^{-3(x-L/x)^2}$, with $A$ and $\sigma$ constants, and
- the solution is to have the fixed value $\varphi = 0$ at $x = 0$ and the fixed value $\varphi = 100$ at $x = L$, i.e., we have a source that is concentrated near the middle of the region of interest and we impose Dirichlet boundary conditions on both ends.

Then, explore the character of the solutions for various values of $\sigma$. Use two-noded elements of equal size. Note: This solution to this problem corresponds physically to the steady state temperature in a one-dimensional rod whose ends are maintained at the fixed temperatures 0 and 100, respectively, and whose middle is heated by a source that provides a constant energy input.

15.14. Repeat the previous exercise but pursue its solution this time by using elements of varying length designed to recognize that, especially if the Gaussian function is sharply peaked in the center, it might be wise to use smaller elements in that region.

15.15. (a) Solve Eq. (15.179) for the constants $a^{(e)}, b^{(e)},$ and $c^{(e)}$ and then verify the expressions given in Eqs. (15.182)-(15.184) for the shape functions. Remember that symbol manipulating programs are available. (b) For each node $i$, verify that the function $N_i^{(e)}(x, y)$ as given by Eqs. (15.182) and (15.184) is zero not only at the two nodes not identified by its index but also along the entire line joining those two nodes. Hint: Set the determinant in the numerator of the expression giving $N_i^{(e)}$ to zero, thereby obtaining the equation of a line in the plane. Verify that this line is, in fact, the line joining the two described nodes. (c) Find the functions to which these functions reduce when $(x_0, y_0) = (-1, 0), (x_1, y_1) = (1, 0), (x_2, y_2) = (0, 1)$ and demonstrate that each is zero where it is supposed to be zero and one where it is supposed to be one.

15.16. Consider a four-node element in two dimensions. Let the nodes be at $(x_0, y_0), (x_1, y_1), (x_2, y_2),$ and $(x_3, y_3)$, and take the interpolating function to be $\tilde{\varphi} = ax + by + cy + dxy$. (a) Find the four shape
functions in general terms, showing that those functions can be expressed in the form

\[ N_0(x, y) = \frac{1}{\Delta} \begin{vmatrix} 1 & x & y & xy \\ 1 & x_1 & y_1 & x_1 y_1 \\ 1 & x_2 & y_2 & x_2 y_2 \\ 1 & x_3 & y_3 & x_3 y_3 \end{vmatrix} \quad \text{where} \quad \Delta = \begin{vmatrix} 1 & x_0 & y_0 & x_0 y_0 \\ 1 & x_1 & y_1 & x_1 y_1 \\ 1 & x_2 & y_2 & x_2 y_2 \\ 1 & x_3 & y_3 & x_3 y_3 \end{vmatrix} \]

and in similar forms for \( N_1(x, y), N_2(x, y), \) and \( N_3(x, y). \) (b) Then, show that, if the nodes lie at the corners of a square of side \( s, \) i.e., the nodes are—in order—at \((0, 0), (s, 0), (s, s), \) and \((0, s), \) the shape functions expressed in terms of the variables \( \xi = x/s \) and \( \eta = y/s \) are

\[ N_0(\xi, \eta) = (1 - \xi)(1 - \eta) \quad N_1(\xi, \eta) = \xi(1 - \eta) \quad N_2(\xi, \eta) = \xi \eta \quad N_3(\xi, \eta) = \eta(1 - \xi) \]

(c) Finally generate surface graphs of the four functions and verify that the shape function associated with each node is zero everywhere along the two edges that intersect at the diagonally opposite node. \( \text{Hints: (1) Remember that symbol manipulating programs are available. (2) Some results along the way to a solution are very involved. Presentation of the intermediate results can certainly be suppressed.} \)

15.17. Find the solution for steady state temperature when a square as in the text has its lower edge maintained at 0, the temperature on its left edge rises linearly from 0 to 50, that on its upper edge rises linearly from 50 to 100, and its right edge is insulated.

15.18. Proof that the full discretization of the wave equation as described towards the end of Section 15.16.1 leads to an unstable method unless \( \alpha \leq 1 \) is extremely difficult. (A proof is worked out in pages 16–29 of \textit{Finite Difference Methods for Partial Differential Equations}, George Forsythe and Wolfgang Wasow (John Wiley and Sons, New York, 1960).) This exercise asks not for that proof but only that you obtain evidence supporting the existence of that instability by recasting \texttt{fdmwave1d} to solve the wave equation subject to the boundary conditions \( u(0, t) = u(t, 0) = 0 \) and the initial conditions \( u(x, 0) = 0, \partial u(x, 0)/\partial x = 0. \) The solution should, of course, be zero at all subsequent times, since we have started the string in its equilibrium position with zero velocity. Now, suppose that a computer roundoff error occurs such that, instead of being zero at all nodes, \( u(x, \Delta t) = 0 \) everywhere except at one node near the middle of the string and, at that node \( u(x, \Delta t) \) mistakenly acquires the value 1 (one). Use your program to solve this problem for choices of the parameters that make \( \alpha = 0.5 \) and solve it again for other choices that make \( \alpha = 1.5. \) Look at the solution with print or plot frequency 1 and compare the way in which that one disruption of a value at the end of the first time step propagates forward in time for the two values of \( \alpha. \)

15.19. Proof that the full discretization of the diffusion equation as described towards the end of Section 15.16.2 leads to an unstable method unless \( \gamma \leq 1/2 \) is extremely difficult. (A proof is worked out in pages 92–98 of \textit{Finite Difference Methods for Partial Differential Equations}, George Forsythe and Wolfgang Wasow (John Wiley and Sons, New York, 1960).) This exercise asks not for that proof but only that you obtain evidence supporting the existence of that instability by recasting \texttt{fdmdiffus1d} to solve the diffusion equation subject to the boundary conditions \( u(0, t) = u(t, 0) = 0 \) and the initial condition \( u(x, 0) = 0. \) The solution should, of course, be zero at all subsequent times, since we have started the temperature distribution with its equilibrium values. Now, suppose that a computer roundoff error occurs such that, instead of being zero at all nodes, \( u(x, \Delta t) = 0 \) everywhere except at one node near the middle of the rod and, at that node \( u(x, \Delta t) \) mistakenly acquires the value 1 (one). Use your program to solve this problem for choices of the parameters that make \( \gamma = 0.25 \) and other choices that make \( \gamma = 1.0. \) Look at the solution with print or plot frequency 1 and compare the way in which that one disruption of a value at the end of the first time step propagates forward in time for the two values of \( \gamma. \)

15.20. The inhomogeneous Helmholtz equation in two-dimensional Cartesian coordinates is

\[ \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + k^2 u = r(x, y) \]

where \( k^2 \) is a constant and \( r(x, y) \) is the inhomogeneity. Apply finite difference methods to show
15.21. Starting with \( v = \sqrt{RT/m_\alpha} \), set \( T = T_0 + \Delta T \) and expand about \( T_0 \) and show that for small variations \( \Delta T \) about this base value, the speed of sound varies linearly with \( \Delta T \).

15.22. In the text, we used the formula in Eq. (15.168) as the basis for an iterative algorithm for solving Laplace’s equation, i.e., for *relaxing* the initial guess to a final solution. This formula estimates the next iterate at a particular node as the average of that node’s four nearest neighbors. An alternative approach, known as *over-relaxation*, includes a contribution from the current value in the node itself, specifically

\[
u_{i,j} = (1 - \alpha) u_{i,j} + \alpha \frac{u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1}}{4}
\]

where \( \alpha = 1 \) reduces this equation to Eq. (15.168) and \( 1 < \alpha < 2 \).\(^{56}\) While the optimal choice of \( \alpha \) depends on the size of the grid used, in general using a value of \( \alpha > 1 \) will speed the convergence of the iterative process to an acceptably accurate solution. Recast \texttt{fdmlap2d} to (1) accept a value of \( \alpha \) as input, (2) implement over-relaxation, (3) monitor the absolute value of the node-by-node change between successive iterates, and (4) print out the maximum value of that change at the end of each iteration. Then, use your program to explore the impact of various values of \( \alpha \) on the rate of convergence for the specific example treated in the text.

15.23. Recast \texttt{fdmlap2d} so that iteration stops when the largest change occurring at any single node from one iterate to the next does not exceed an externally prescribed tolerance. To avoid all possibility of an infinite loop, you should halt iteration *either* when the tolerance has been reached (or exceeded) *or* when some prescribed number of iterations has taken place. Code so that your program displays the actual tolerance achieved. Further, if execution terminates because the prescribed tolerance is *not* achieved, your program should print a message that alerts you to the fact that the prescribed tolerance was not achieved. Test your program with the same example as was used in the text. *Hints:* Before starting an iteration, set a variable, say \texttt{maxch}, equal to zero. As you calculate a new value for each node in the iteration, store the result in a temporary variable so you can compare that value with the old value it will replace, updating \texttt{maxch} to the absolute value of the difference between the new and the old values *if and only if* that difference exceeds the difference already stored in \texttt{maxch}. Then, substitute the new value for the old in the array containing the evolving solution and go on to the next node. Once the iteration is completed, \texttt{maxch} will contain the absolute value of the largest change at any node during that single iteration. If \texttt{maxch} is less than the prespecified convergence criterion, stop the iteration; otherwise conduct one more iteration.

15.24. Consider the problem defined by the two-dimensional Poisson equation

\[
\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = -(1 - x^2)(1 - y^2)
\]

to be solved in the square region \( R \) defined by \( -1 \leq x, y \leq 1 \) subject to the Dirichlet boundary conditions requiring \( U = 0 \) on the entire boundary of \( R \). Assume all variables are dimensionless. (a) Obtain a surface plot of this inhomogeneity over the \( x, y \) plane. (b) Recasting \texttt{fem2d} to incorporate the given boundary conditions and to address the non-zero inhomogeneity find, explore, and (using surface and contour plots as you deem appropriate) display the solution \( U(x, y) \) in the square region \( R \). Compare the resulting equation with Eq. (15.69) as a guide to interpreting the present problem as a modification of the problem addressed by \texttt{fem2d}. (c) Find the two-dimensional vector field \( \mathbf{V} = -\nabla U = -\nabla U \) in \( R \) and display that field graphically in whatever ways seem appropriate.

\(^{56}\) \( \alpha \geq 2 \) generates an unstable algorithm.
15.25. (a) Starting with the Taylor series

\[ \varphi(x + \Delta x) = \varphi(x) + \Delta x \varphi'(x) + O(\Delta x^2) \]

evaluate

\[ \varphi(x + \Delta x) - \varphi(x) \]

to show that

\[ \varphi'(x) = \frac{\varphi(x + \Delta x) - \varphi(x)}{\Delta x} + O(\Delta x) \]

That is, derive the forward difference formula in Eq. (15.74). (b) Deduce the more accurate forward difference formula

\[ \varphi'(x) = -\frac{\varphi(x + 2 \Delta x) + 4\varphi(x + \Delta x) - 3\varphi(x)}{2 \Delta x} + O(\Delta x^2) \]

Hint: Start by using the Taylor expansion

\[ f(x + h) = f(x) + h f'(x) + \frac{1}{2} h^2 f''(x) + O(h^3) \]

to expand \( \varphi(x + \Delta x) \) and \( \varphi(x + 2 \Delta x) \).

15.26. Starting with Eqs. (15.85) and (15.86), derive Eqs. (15.87), (15.88), and (15.89).

15.27. Following the pattern illustrated in Section 15.16.1, develop equations corresponding to Eqs. (15.145)–(15.148) to solve the wave equation

\[ \frac{1}{c^2} \frac{\partial^2 u(x, t)}{\partial t^2} = \frac{\partial^2 u(x, t)}{\partial x^2} \]

subject to the boundary and initial conditions

\[ u(0, t) = 0 \quad ; \quad u(L, t) = 0 \quad ; \quad u(x, 0) = f(x) \quad ; \quad \frac{\partial u(x, t)}{\partial t} \bigg|_{t=0} = 0 \]

These equations describe a string of length \( L \) that is fixed at both ends and set into motion by drawing it aside to the initial shape given by \( f(x) \) and released from rest. Of course, for consistency, \( f(0) \) and \( f(L) \) must both be zero, but \( f(x) \) is otherwise unconstrained. Then, develop a program analogous to \texttt{fdmwave1d} to solve this problem, testing your program with

\[ f(x) = A \sin \frac{n \pi x}{L} \]

for \( n = 1, 2, \) and \( 3 \). You should recognize that this exercise asks you to find the lowest three normal modes in the vibrations of this string. You should, of course, expect the results of your program to reveal what you know to be the periodic motion in these three modes of oscillation of the string. Your program should also allow you to confirm that the frequencies of the three modes \( f_1, f_2, \) and \( f_3 \) are related by \( f_2 = 2f_1 \) and \( f_3 = 3f_1 \). Agreement with these expectations provides some level of confidence in the adequacy of your choice of parameters.
15.A  Program(s) for FDM Approach to 1D Problem

15.A.3  Listing of fdm1d.m (OCTAVE)

% ***** Command file fdm1d.m *****
% ***** Note that, when this command file has completed executing,
% all variables to which it assigns values---and in
% particular x and phi---will be accessible at OCTAVE's main
% command level.
% ***** Request input of necessary parameters.

N = input( 'Enter number of segments (N): ' );
alpha = input( 'Enter alpha: ' );
beta = input( 'Enter beta: ' );
f = input( 'Enter f: ' );
L = input( 'Enter L: ' );
p = input( 'Enter p: ' );
gamma = input( 'Enter gamma: ' );
q = input( 'Enter q: ' );

% ***** Calculate segment size, square of segment size, and
% coordinates of nodes.

dx = L/N; dx2 = dx^2;
x = dx*[ 0 : N ];

% ***** Determine vector of inhomogenieties and coefficient matrix.

inhomo = zeros(N+1,1) + 1.0;
inhomo = f*dx2*inhomo;
inhomo(1) = p;
inhomo(N+1) = inhomo(N+1) + 2.0*q*dx;

for i=1:N+1 cf(i,i) = 2.0*alpha + beta*dx2; end
for i=1:N cf(i,i+1) = -alpha; end
for i=1:N cf(i+1,i) = -alpha; end
phi = cf\inhomo;
15.B Evaluating Integrals for 1D FEM Problem

15.B.3 ... using Mathematica

An appropriate batch file to use Mathematica to evaluate the important integrals in Section 15.1.7 is named FEM1DCalcs.mtm, can be copied from the directory $HEAD/mathematica, and invoked within Mathematica with the statement Get[ "FEM1DCalcs.mtm"] issued at Mathematica’s prompt for commands. The batch file is listed (without output) in the remainder of this subsection.

(* FEM1DCalcs.mtm *)

(* This batch file evaluates the several integrals that appear in *)
(* setting up coding to solve the 1D FEM problem in CPSUP. *)

(* Create interpolation functions and their derivatives *)

n[1] = (x2-x)/l; n[2] = (x-x1)/l ;
dnx[1] = D[ n[1], x ];
dnx[2] = D[ n[2], x ];

(* Create integrand for K, then integrate and display *)

K1 = Table[ dnx[i]*dnx[j], {i, 2}, {j, 2} ];
K2 = Table[ n[i]*n[j], {i, 2}, {j, 2} ];
K = a*K1 + b*K2;
K = Integrate[ K, {x, x1, x2} ];
K = Simplify[ K /. x2 -> x1 + l ];
Print["K =", K // MatrixForm ]

(* Evaluate and display b *)

b = Integrate[ f*n[1]*n[2], {x, x1, x2} ];
b = Simplify[ b /. x2 -> x1 + l ];
Print["b = ", b ]

15.C Program(s) for FEM Approach to 1D Problem

15.C.3 Listing of fem1d.m (OCTAVE)

# ****** Command file fem1d.m ******

# ****** Note that, when this command file has completed executing, #
# all variables to which it assigns values---and in #
# particular x and phi---will be accessible at OCTAVE’s main #
# command level.
#
# BLOCK 1 Request input of necessary parameters and assure that each #
# is stored with the proper data type. Add calculation of the #
# length of the rod and values of x along the rod.

M = input( 'Enter number of segments (M): ' );
alpha = input( 'Enter alpha: ' );
beta = input( 'Enter beta: ' );
f = input( 'Enter f: ' );
l = input( 'Enter l: ' );
p = input( 'Enter p: ' );
gamma = input( 'Enter gamma: ' );
q = input( 'Enter q: ' );
length = l*M;
x = l * [ 0 : M ];

# BLOCK 2 Determine coefficient (stiffness) matrix before incorporation of boundary conditions.
K = zeros(M+1,M+1); # Create (M+1)x(M+1) array of zeros
S = alpha/l + beta*l/3.0; # Evaluate common quantities
S2 = 2.0*S;
T = -alpha/l + beta*l/6.0;
K(1,1) = S; # Set diagonal elements of K
for i = 2:M K(i,i) = S2; end
K(M+1,M+1) = S;
for i = 1:M K(i+1,i) = T; end # Set elements above and below main diagonal of K
for i = 1:M K(i,i+1) = T; end

# BLOCK 3 Create vector of inhomogenieties.
b = zeros(M+1,1); # Create M+1 element vector of zeros
U = f*l/2.0; # Evaluate common quantities
U2 = 2.0*U;
b(1) = U; # Set elements of b
for i = 2:M b(i) = U2; end
b(M+1) = U;

# BLOCK 4 Incorporate mixed boundary condition
K(M+1,M+1) = K(M+1,M+1) + gamma;
b(M+1) = b(M+1) + q;

# BLOCK 5 Incorporate Dirichlet boundary conditions.
K(1,1) = 1.0;
b(1) = p;
for i = 2:M+1 K(i,1) = 0.0; end
for i = 2:M+1 b(i) = b(i) - K(i,i)*p; end
for j = 2:M+1 K(1,j) = 0.0; end

# BLOCK 6 Solve system using LU decomposition.
phi = K\b;
15.D OCTAVE Programs for FDM Approach to 2D Problems

15.D.1 Listing of fdmwave1d.m

% ***** Command file fdmwave1d.m *****

% ***** Request input of necessary parameters and assure that
% each is stored with the proper data type.

N = input( 'Enter number of segments (N): ' ); N = fix(N);
dt = input( 'Enter time step (dt): ' );
T = input( 'Enter number of time steps (T): ' ); T = fix(T);
c = input( 'Enter speed of propagation (c): ' );
L = input( 'Enter length of string (L): ' );
delay = input( 'Delay between plots in secs (delay): ' );
f = input( 'Plot frequency (f): ' ); f = fix(f);

% ***** Determine segment length, set independent variables,
% and evaluate and display parameter; test for stability.

dx = L/N;
x = dx*[ 0 : N ];
alpha = c^2*dt^2/dx^2;
alpha
if alpha > 1.0
    disp( 'Error: alpha > 1; execution halted' );
    return;
end

% ***** Define vectors for past, current, and new solutions

u1 = zeros( N+1, 1 );     % For past solution
u2 = zeros( N+1, 1 );     % For current solution
u3 = zeros( N+1, 1 );     % For next solution

% ***** Set and display initial displacement

b = 8.0*pi/L;
for i=fix(3.0*N/8.0)+1:fix(5.0*N/8.0)
    u1(i) = 1.0 + cos(b*(x(i)-L/2.0));
end
plot( x, u1, 'linewidth', 3, 'color', 'black' )
set(gca, 'ylim', [-2.0,2.0] );
pause( delay );

% ***** Set solution at time dt with initial velocity
% equal to zero; display if requested

u2 = u1;
if f == 1
    plot( x, u2, 'linewidth', 3, 'color', 'black' )
end
15.D. OCTAVE PROGRAMS FOR FDM APPROACH TO 2D PROBLEMS

```octave
set(gca, 'ylim', [-2.0,2.0] );
pause( delay );
end
%
% ***** Calculate solution repeatedly, displaying every f-th result
for j=2:T
  u3(1) = 0;
  for i=2:N
    u3(i) = alpha*u2(i+1) + ...  
      2.0*(1.0-alpha)*u2(i) + alpha*u2(i-1) - u1(i);
  end
  u3(N+1) = 2*alpha*u2(N)+2*(1.0-alpha)*u2(N+1) - u1(N+1);
  u1 = u2; u2 = u3;
  if f*fix(j/f) == j
    plot( x, u2, 'linewidth', 3, 'color', 'black' );
    set(gca, 'ylim', [-2.0,2.0] );
    pause( delay );
  end
end
```

15.D.2 Listing of fdmdiffus1d.m

```octave
% ***** Command file fdmdiffus1d.m *****
% ***** Request input of necessary parameters and assure that each is stored with the proper data type.
N = input( 'Enter number of segments (N): ' ); N = fix(N);
dt = input( 'Enter time step (dt): ' );
T = input( 'Enter number of time steps (T): ' ); T = fix(T);
alpha = input( 'Enter value of alpha (alpha): ' );
L = input( 'Enter length of rod (L): ' );
delay = input( 'Delay between plots in secs (delay): ' );
f = input( 'Plot frequency (f): ' ); f = fix(f);

% ***** Determine segment length, set independent variables, and evaluate and display parameter; test for stability.
dx = L/N;
x = dx*[ 0 : N ];
gamma = alpha^2*dt/dx^2;
gamma
if gamma > 0.5
  disp( 'Error: gamma > 0.5; execution halted' );
  return;
end
% ***** Define vectors for current and new solutions
```
u1 = zeros(N+1, 1); % For current solution
u2 = zeros(N+1, 1); % For next solution

% ***** Set and display initial temperature distribution

b = 8.0*pi/L;
for i=fix(3*N/8)+1:fix(5*N/8)
    u1(i) = 1.0 + cos(b*(x(i)-L/2.0));
end

plot(x, u1, 'linewidth', 3, 'color', 'black')
title('t = 0 s', 'fontsize',20)
set(gca, 'YLim', [ 0.0, 2.0 ], 'fontsize',16)

% ***** Calculate solution repeatedly, displaying every % f-th result.

for j=1:T
    u2(1) = 0.0;
    for i=2:N
        u2(i) = gamma*u1(i-1) + (1.0-2.0*gamma)*u1(i) + gamma*u1(i+1);
    end
    u2(N+1) = 2*gamma*u1(N) + (1-2*gamma)*u1(N+1);
    if f*fix(j/f) == j
        plot(x, u2, 'linewidth', 3, 'color', 'black')
title(['t = ', num2str(dt*j), ' s'], 'fontsize',20)
set(gca, 'YLim', [ 0.0, 2.0 ], 'fontsize',16)
pause(delay);
    end
    u1 = u2;
end

15.D.3 Listing of fdmlap2d.m

# ***** Command file fdmlap2d.m *****

# ***** Acquire controlling parameters *****

N = input('Enter number of segments (N): '); N = fix(N);
L = input('Enter length of side(L): ');
maxits = input('Maximum number of iterations (maxits): ');
maxits = fix(maxits);
f = input('Display frequency (f): ');
f = fix(f);

# ***** Calculate grid spacing, values of x and y at grid points *****

dx = L/N;
x = dx*[0 : N]; y = x;

# ***** Create and initialize array for solution *****

u = zeros(N+1, N+1);
u(1,:) = 100.0;
u(:,1) = 100.0 - 100.0*x/L;

# ***** Display solution on the screen *****

printf( '\nIteration 0\n' )

u

# ***** Iterate the specified number of times, displaying
# every f-th iterate on the screen *****

for itcnt=1:maxits
    for i = 2:N
        for j = 2:N
            u(i,j) = 0.25*(u(i+1,j) + u(i-1,j) + u(i,j+1) + u(i,j-1));
        end
        u(i,N+1) = 0.25*(2.0*u(i,N) + u(i-1,N+1) + u(i+1,N+1));
    end
    if f*fix(itcnt/f) == itcnt
        printf( '\nIteration #d\n', itcnt )
        u
    end
end

15.D.4 Listing of fdmlap2d_plot.m

# ***** Command file fdmlap2d_plot.m *****

# ***** Acquire controlling parameters *****

N = input( 'Enter number of segments (N): ' ); N = fix(N);
L = input( 'Enter length of side(L): ' );
maxits = input( 'Maximum number of iterations (maxits): ' );
maxits = fix( maxits );
f = input( 'Display frequency (f): ' ); f = fix(f);
delay = input( 'Delay between plots in secs (delay): ' );

# ***** Calculate grid spacing, values of x and y at grid points *****

dx = L/N;
x = dx*[0:N]; y = x;

# ***** Create and initialize array for solution *****

u = zeros(N+1, N+1);
u(1,:) = 100.0;
u(:,1) = 100.0 - 100.0*x/L;

# ***** Create graphical display of solution on the screen *****

mesh( x,y,flip(u,1), 'edgecolor', 'black' )
xlabel('x','fontsize',16); ylabel('y','fontsize',16);
zlabel('u','fontsize',16);
CHAPTER 15. SOLVING PARTIAL DIFFERENTIAL EQUATIONS

set(gca,'fontsize',12)
view( [40 10] )
pause( delay )

# ***** Iterate the specified number of times, displaying
# every f-th iterate on the screen *****

for itcnt=1:maxits
    for i = 2:N
        for j = 2:N
            u(i,j) = 0.25*(u(i+1,j) + u(i-1,j) + u(i,j+1) + u(i,j-1));
        end
        u(i,N+1) = 0.25*(2.0*u(i,N) + u(i-1,N+1) + u(i+1,N+1));
    end
    if f*fix(itcnt/f) == itcnt
        mesh(x,y,flip(u,1), 'edgecolor', 'black' )
        xlabel('x','fontsize',16); ylabel('y','fontsize',16);
        zlabel('u','fontsize',16);
        set(gca,'fontsize',12)
        view( [40 10] )
        pause( delay )
    end
end

15.E A Useful Integral

In this appendix, we deduce a two-dimensional analog to the familiar one-dimensional formula for integration by parts. Consider the integral

\[ I = \int \int_{\Omega} \psi \left( \frac{\partial V_x}{\partial x} + \frac{\partial V_y}{\partial y} \right) \, dx \, dy \quad (15.225) \]

where \( \psi, V_x, \) and \( V_y \) are functions of \( x \) and \( y \), and \( \Omega \) is the region in the \( xy \) plane over which the 2D integral is to be evaluated. Recast the integral in the form

\[ I = \int \int_{\Omega} \left( \nabla \times Q \right) \cdot \hat{k} \, dx \, dy - \int \int_{\Omega} \left( V_x \frac{\partial \psi}{\partial x} + V_y \frac{\partial \psi}{\partial y} \right) \, dx \, dy \quad (15.226) \]

Now, introduce the vector \( Q \) whose \( x \) and \( y \) components are \( -\psi V_y \) and \( \psi V_x \), respectively. Since

\[ \frac{\partial (\psi V_x)}{\partial x} + \frac{\partial (\psi V_y)}{\partial y} = \frac{\partial Q_y}{\partial x} - \frac{\partial Q_x}{\partial y} = (\nabla \times Q)_z = (\nabla \times Q) \cdot \hat{k} \quad (15.227) \]

the integral of concern then can be written in the form

\[ I = \int \int_{\Omega} (\nabla \times Q) \cdot \hat{k} \, dx \, dy - \int \int_{\Omega} \left( V_x \frac{\partial \psi}{\partial x} + V_y \frac{\partial \psi}{\partial y} \right) \, dx \, dy \quad (15.228) \]

which, upon invoking Stokes' theorem, we can rewrite further in the form

\[ I = \oint \mathbf{Q} \cdot \mathbf{t} \, dl - \int \int_{\Omega} \left( V_x \frac{\partial \psi}{\partial x} + V_y \frac{\partial \psi}{\partial y} \right) \, dx \, dy \quad (15.229) \]
where \( \Gamma \) is the path in the \( xy \) plane bounding the region \( \Omega \) and \( \hat{t} \) is a unit vector tangent to that path and pointing in the direction of the thumb of the right hand when the fingers grasp the path while piercing the region \( \Omega \) in the \( z \) direction. The one-dimensional integral in this result is the analog of the unintegrated term in the more familiar one-dimensional formula for integration by parts.

That one dimensional integral involves the component of the vector \( \mathbf{Q} \) tangent to the path \( \Gamma \). We are actually better served by casting things in terms of components normal to that curve. Thus, we introduce the unit vector \( \hat{n} \) normal to the curve and pointing outward from the perspective of a viewer in the region \( \Omega \). Since \( \hat{t} = \hat{k} \times \hat{n} \), we find that

\[
\hat{t} = \hat{k} \times \hat{n} = \hat{k} \times (n_x \hat{i} + n_y \hat{j}) = -n_y \hat{i} + n_x \hat{j}
\]

Thus,

\[
\mathbf{Q} \cdot \hat{t} = n_x Q_y - n_y Q_x = n_x \psi V_x + n_y \psi V_y = (\psi V_x \hat{i} + \psi V_y \hat{j}) \cdot \hat{n} = \psi (V_x \hat{i} + V_y \hat{j}) \cdot \hat{n}
\]

In other words, the tangential component of \( \mathbf{Q} \) can be recast as the normal component of the vector from which \( \mathbf{Q} \) was originally derived. Thus, we find that the original integral in Eq. (15.225) can alternatively be evaluated as

\[
I = -\int \int_{\Omega} \left( V_x \frac{\partial \psi}{\partial x} + V_y \frac{\partial \psi}{\partial y} \right) \, dx \, dy + \oint_{\Gamma} (\psi V_x \hat{i} + \psi V_y \hat{j}) \cdot \hat{n} \, dl
\]

### 15.F Evaluating Integrals for 2D FEM Problem

A symbol manipulating program can usefully be invoked to evaluate the several integrals that appear in the solution of the example problem introduced in Section 15.23.

#### 15.F.3 . . . using *Mathematica*

An appropriate batch file to use *Mathematica* to evaluate the necessary integrals when array indices start at 0 is named `FEM2DCalcs0.mtm`, can be copied from the directory `$HEAD/mathematica`, and invoked within *Mathematica* with the statement `Get[ "FEM2DCalcs0.mtm" ]` issued at *Mathematica*’s prompt for commands. The batch file includes the statements

(* FEM2DCalcs0.mtm *)

(* This batch file evaluates the several integrals that appear in *)

(* setting up coding to solve the 2D FEM problem in CPSUP when *)

(* array indices start at 0. *)

(* Create interpolation functions and their derivatives *)

\[
r = \{ 1, x, y \} \\
r0 = \{ 1, x0, y0 \} \\
r1 = \{ 1, x1, y1 \} \\
r2 = \{ 1, x2, y2 \} \\
\]

\[
\text{CapDelta} = \text{Det}[ \{r0, r1, r2\} ]
\]

\[
n0 = \text{Det}[ \{r, r1, r2\} ] / \text{CapDelta} (* Evaluate functions *) \\
n1 = \text{Det}[ \{r0, r, r2\} ] / \text{CapDelta} \\
n2 = \text{Det}[ \{r0, r1, r\} ] / \text{CapDelta}
\]
\[ n = \{ n_0, n_1, n_2 \} \]

(* Simplify functions and evaluate derivatives *)
(* for odd-numbered elements *)

\[ sb = \{ x_1 \rightarrow x_0 + dx, x_2 \rightarrow x_0 + dx, y_1 \rightarrow y_0, y_2 \rightarrow y_0 + dx \} \]
\[ n_{odd} = Simplify[ n /. sb ] \]
\[ dnoddx = D[ n_{odd}, x ] \]
\[ dnoddy = D[ n_{odd}, y ] \]

(* Simplify functions and evaluate derivatives *)
(* for even-numbered elements *)

\[ sb = \{ x_1 \rightarrow x_0 + dx, x_2 \rightarrow x_0, y_1 \rightarrow y_0 + dx, y_2 \rightarrow y_0 + dx \} \]
\[ n_{even} = Simplify[ n /. sb ] \]
\[ dnevenx = D[ n_{even}, x ] \]
\[ dneveny = D[ n_{even}, y ] \]

(* Create integrand for \( K_{odd} \), then integrate *)

\[ K_{odd1} = Table[ dnoddx[[i]]*dnoddx[[j]], \{ i, 3 \}, \{ j, 3 \} ] \]
\[ K_{odd2} = Table[ dnoddy[[i]]*dnoddy[[j]], \{ i, 3 \}, \{ j, 3 \} ] \]
\[ K_{odd3} = Table[ n_{odd}[[i]]*n_{odd}[[j]], \{ i, 3 \}, \{ j, 3 \} ] \]
\[ K_{odd} = ax*K_{odd1} + ay*K_{odd2} + b*K_{odd3} \]
\[ K_{odd} = Integrate[ K_{odd}, \{ y, y_0, y_0+x-x_0 \} ] \]
\[ K_{odd} = Integrate[ K_{odd}, \{ x, x_0, x_0+dx \} ] \]
\[ K_{odd} = Simplify[ K_{odd} ] \]
Print[ "Kodd =", K_{odd}//MatrixForm ]

(* Create integrand for \( K_{odd} \), then integrate *)

\[ K_{even1} = Table[ dnevenx[[i]]*dnevenx[[j]], \{ i, 3 \}, \{ j, 3 \} ] \]
\[ K_{even2} = Table[ dneveny[[i]]*dneveny[[j]], \{ i, 3 \}, \{ j, 3 \} ] \]
\[ K_{even3} = Table[ n_{even}[[i]]*n_{even}[[j]], \{ i, 3 \}, \{ j, 3 \} ] \]
\[ K_{even} = ax*K_{even1} + ay*K_{even2} + b*K_{even3} \]
\[ K_{even} = Integrate[ K_{even}, \{ y, y_0+x-x_0, y_0+dx \} ] \]
\[ K_{even} = Integrate[ K_{even}, \{ x, x_0, x_0+dx \} ] \]
\[ K_{even} = Simplify[ K_{even} ] \]
Print[ "Keven =", K_{even}//MatrixForm ]

(* Evaluate \( b_{odd} \) and \( b_{even} \) *)

\[ b_{odd} = Integrate[ f*n_{odd}, \{ y, y_0, y_0+x-x_0 \} ] \]
\[ b_{odd} = Integrate[ b_{odd}, \{ x, x_0, x_0+dx \} ] \]
\[ b_{odd} = Simplify[ b_{odd} ] \]
Print[ "bodd = ", b_{odd} ]

\[ b_{even} = Integrate[ f*n_{even}, \{ y, y_0+x-x_0, y_0+dx \} ] \]
\[ b_{even} = Integrate[ b_{even}, \{ x, x_0, x_0+dx \} ] \]
\[ b_{even} = Simplify[ b_{even} ] \]
Print[ "beven = ", b_{even} ]
15.F. EVALUATING INTEGRALS FOR 2D FEM PROBLEM

(* Evaluate needed g's *)

\[ g = \text{nodd} /\ .\ x \rightarrow x0+dx \]
\[ g = \text{Integrate}[ q*g, \{y, y0, y0+dx\} ] \]
\[ \text{Print}[ "g = ", \ g \] \]

An appropriate batch file to use *Mathematica* to evaluate the necessary integrals when array indices start at 1 is named *FEM2DCalcs1.mpl*, can be copied from the directory $\text{$HEAD/mathematica}, and invoked within *Mathematica* with the statement \text{batch} ( "FEM2DCalcs1.mtm" ) issued at *Mathematica*’s prompt for commands. This batch file includes the statements
# ***** Command file fem2d.m *****

# ***** Enter parameters *****

L = input('Enter length of side (L): ');
d = input('Enter number of segments (d): '); d = fix(d);
alpha_x = input('Enter alpha_x: ');
alpha_y = input('Enter alpha_y: ');
beta = input('Enter beta: ');
f = input('Enter f: ');
p1 = input('Enter value for top edge (p1): ');
p2 = input('Enter value for bottom edge (p2): ');
q = input('Enter q: ');

# ***** Find number of elements M, nodes N, and segment side d *****

M = 2*d^2; # Calculate number of elements
N = (d+1)^2; # Calculate number of nodes
dx = L/d; # Calculate segment size

# ***** Find coordinates of nodes *****

x = zeros(1,N); # Create array to store x values
y = zeros(1,N); # Create array to store y values
ct = 0; # Initialize a counter variable
for i = 0:d # Start row number loop
    for j = 0:d # Start column number loop
        ct=ct+1; # Increment counter
        x(ct) = i*dx; # Find x coordinate
        y(ct) = L - j*dx; # Find y coordinate
    end # End loops
end

# ***** Create connectivity matrix *****

cm = zeros(3,M); # Create 3 by M null array
for e = 1:M # Loop through all elements
    vl = fix((e-1)/(2*d)) + 1; # Find index of vertical line
    cm(1,e) = fix((e-1)/2) + 1 + vl; # Find global number of node 1
    if 2*fix(e/2) == e # If e is even
        cm(2,e) = cm(1,e) + d + 1; # find global number of node 1
        cm(3,e) = cm(1,e) + d; # find global number of node 2
    else # If e is odd
        cm(2,e) = cm(1,e) + d; # find global number of node 1
        cm(3,e) = cm(1,e) - 1; # find global number of node 2
    end
end
# ***** Find [K] for odd and even numbered elements *****

\begin{verbatim}
Kodd = zeros(3,3);  # Create two 3 by 3 arrays
Keven = zeros(3,3);

bx = beta*dx^2;  # Evaluate a common quantity

Keven(1,1) = (bx + 6*alpha_x)/12;  # Assign the appropriate value to
Keven(2,2) = (bx + 6*alpha_x + 6*alpha_y)/12;  # each K(i,j). Note that the array
Keven(3,3) = (bx + 6*alpha_y)/12;  # are symmetric, and that Kodd
Keven(1,2) = (bx - 12*alpha_x)/24;  # includes all of the same values as
Keven(2,1) = Keven(1,2);  # Keven, but in different locations.
Keven(2,3) = (bx - 12*alpha_y)/24;
Keven(3,2) = Keven(2,3);
Keven(1,3) = bx/24;
Keven(3,1) = bx/24;

Kodd(1,1) = Keven(3,3);
Kodd(2,2) = Keven(1,1);
Kodd(3,3) = Keven(2,2);
Kodd(1,2) = Keven(1,3);
Kodd(2,1) = Kodd(1,2);
Kodd(2,3) = Keven(1,2);
Kodd(3,2) = Kodd(2,3);
Kodd(1,3) = Keven(2,3);
Kodd(3,1) = Kodd(1,3);
\end{verbatim}

# ***** Create stiffness matrix and vector \{b\} *****

\begin{verbatim}
K = zeros(N,N);  # Create arrays to store values of
b = zeros(N,1);  # K(i,j) and b(i)

for e = 1:M  # Count through element numbers
    for i = 1:3  # For each local node of an element
        b(cm(i,e)) = b(cm(i,e)) + f*dx^2/6;  # place its contributions at the
    for j = 1:3  # correct locations in [K] and \{b\}
        if 2*fix(e/2) ~= e
            K(cm(i,e),cm(j,e)) = K(cm(i,e),cm(j,e)) + Kodd(i,j);
        else
            K(cm(i,e),cm(j,e)) = K(cm(i,e),cm(j,e)) + Keven(i,j);
        end
    end
end

# ***** Incorporate Dirichlet boundary conditions *****

for i = 0:d
    u = i + 1;  # Nodes on the left boundary
    s = i*(d+1) + 1;  # Nodes on the top boundary
    t = (i+1)*(d+1);  # Nodes on the bottom boundary
    p3 = (p1-p2)/L * y(u) + p2;  # Find values of \phi on left boundary
    for j = 1:N
\end{verbatim}
K(j,s) = 0; # Set rows in K to zero where value
K(j,t) = 0; # of phi is known
K(j,u) = 0;
b(s) = p1; # Set values of b
b(t) = p2;
b(u) = p3;
if j ~= s # Reflect influence
    b(j) = b(j) - K(s,j)*p1; # of known values of
end # phi in the other
K(s,j) = 0; # equations
if j ~= t
    b(j) = b(j) - K(t,j)*p2;
end
K(t,j) = 0;
if j ~= u
    b(j) = b(j) - K(u,j)*p3;
end
K(u,j) = 0;
end
K(s,s) = 1; # Set the appropriate entry to 1 in
K(t,t) = 1; # the rows where phi is known
K(u,u) = 1;
end

# ***** Incorporate Neumann boundary condition *****

g = zeros(N,1); # Create vector to store g values
for i = 1:d-1
    nd = d*(d+1)+i+1; # Nodes on right boundary
    g(nd) = q * dx;
end
b = b+g; # Store the vector {b+g} in {b}

# ***** Solve the system of equations *****

phi = K\b; # Solve the equation K*phi = b
# writing the results to phi

# ***** Store solution and node locations in an array *****

A = zeros(d+1,d+1); # Create a d+1 by d+1 array
cnt=0; # Initialize a counter
for i = 1:d+1 # Use nested for loops to write all
    for j = d+1:-1:1 # entries in A
        cnt=cnt+1; # Increment counter
        A(i,j) = phi(cnt);
    end # End inner loop
end # End outer loop
Appendix A

Introduction to \LaTeX

Note: All \LaTeX\ source files (.tex, .template), all Windows batch files (.bat) program (.pro) files, and all UNIX command files (no file type) referred to in this chapter are available in the directory $\texttt{SHEAD/tex}$, where (as defined in the Local Guide) $\texttt{SHEAD}$ must be replaced by the appropriate path for your site.

During the 1970s, when Donald Knuth (a Stanford University computer scientist) was creating his monumental, several-volume series of books on all aspects of computer science, he recognized the need for a computer-based type-setting/word-processing system tailored to the needs of authors of technical manuscripts. Interrupting his main project, he developed \TeX\ as the essential engine to support computerized technical type setting. The first version of \TeX\ was made available to the using public in 1978 and a revised (and final) version was published in 1984. Written by Leslie Lamport, \LaTeX\ is a versatile and extensive collection of macros that sit on top of \TeX\ and facilitate exploitation of \TeX\’s capabilities. Broadly, \LaTeX\ is a document preparation system that formats equations, tables, and illustrations as easily as plain text. That the whole complex (\TeX, \LaTeX, and many other components) has been deliberately placed in the public domain, that carefully tested versions exist for essentially every computing platform and operating system, and that many scientific journals and an increasing number of publishers will accept manuscripts submitted as \LaTeX\ files together provide substantial incentive for learning to use this tool. Although it is not \TeX\y, it has powerful and sophisticated capabilities, most of which are not described in this Appendix.

We convey only the basics with, however, the hope that we will cause you to realize that essentially any desired formatting at all should be possible if only you can figure out how to achieve it. For further information, try surfing the web in your favorite browser and looking at the \LaTeX\ User’s Guide and Reference Manual (which we will refer to as The \LaTeX\ Manual), the \LaTeX\ Companion, and/or to any of the several other \TeX\ and \LaTeX\ books. The wisdom of taking fifteen minutes every now and then to browse in these publications cannot be overstressed; as you become more skilled, the fine print and other subtleties will gradually be more and more meaningful.

---

1 As per Knuth’s instruction (see the first two paragraphs of Chapter 1 of The \TeX\book as identified in Section A.18), \TeX\ is pronounced ‘tech’ as in ‘technology’.
2 Here, individuals disagree on the pronunciation of the ‘la’, and Leslie Lamport offers no guidance. Some say la\TeX\ while others say lay\TeX, with the emphasis on the first syllable in both cases. We simply must become accustomed to each other’s preferences.
3 The primary site for information (history, current plans, downloads, ...) for \TeX, \LaTeX, and numerous other publicly available components of \TeX\ and its derivatives is the web site of the \TeX\ Users’ Group (TUG), \url{www.tug.org}. This organization maintains CTAN (the Comprehensive \TeX\ Archive Network), which has a handful of backbone machines around the world and a number of mirror sites, from any of which an enormous number of files associated with the \TeX/\LaTeX\ system can be downloaded.
4 What you see is what you get.
5 See Section A.18 for more detailed references.
6 You may also find the web pages of TUG at the URL \url{http://www.tug.org} to be valuable. One link on that page points to an engaging description of the history of the development of \TeX. 

459
A.1 Creating a Simple Document

Producing a final document with \LaTeX involves a number of steps, beginning with the creation of an ASCII text file containing the \LaTeX “source code”—hereafter simply “code”—for the document. Appropriate conversion programs are then invoked to create either a PostScript or a PDF file containing the formatted document.\footnote{Other formats (HTML, ePub, …) may also be produced, though this Appendix will not address those alternatives.} Finally, the PostScript or PDF file will be viewed on the screen or sent to a printer. In this section, we describe how to structure the code for a very simple document and then how to carry out the remaining steps to convert that code into a displayed or printed document. The remainder of this Appendix will explain numerous embellishments that might be invoked in the code. Processing that code to produce the final document is (more or less) independent of the complexity of the code.

To keep the discussion comparatively simple, we assume initially that your document incorporates no figures, does not have internal cross references, and includes neither a table of contents or an index. Those embellishments will be described in later sections.

A.1.1 Structuring a \LaTeX Source File

As input, \LaTeX requires an ASCII file, which can be created with any text editor (\texttt{xemacs}, \texttt{notepad}, \texttt{wordpad}, \texttt{winedt}, \texttt{gedit}, …) and which contains both the text of the desired document and embedded formatting commands. All of \LaTeX’s commands begin with a backslash (\). \LaTeX distinguishes upper and lower case letters; while most standard commands and parameters use exclusively lower case letters, we must nonetheless pay attention to case.

Three commands are sufficient to create the simplest document containing straight text formatted with \LaTeX’s defaults. The very first required command in the code for \textit{any} \LaTeX document specifies the document class and has the form\footnote{Be aware that some of the features described in this Appendix are specific to \LaTeXX and will not work with the previous version (\LaTeX 2.09). \LaTeXX, however, will automatically enter 2.09 emulation mode if the first line in the code is the now obsolete command \texttt{documentstyle} instead of the new command \texttt{documentclass}.}

\begin{verbatim}
\documentclass{Class} \ or \ \documentclass[options]{Class}
\end{verbatim}

where

- \textit{Class} is any one of the options shown in Table A.1 and
- \textit{options} may be omitted altogether if the defaults are acceptable. The most commonly used options, one or more of which may be separated by commas in a string containing no spaces, are listed in Table A.2.

The document itself must be enclosed between the second and third of the essential commands, the second marking the beginning of the document proper and having the form

\begin{verbatim}
\begin{document}
\end{verbatim}

and the third marking the end of the document and having the form

\begin{verbatim}
\end{document}
\end{verbatim}
A.1. CREATING A SIMPLE DOCUMENT

Table A.1: Document classes valid in standard $\LaTeX$. Other available classes may be described in the *Local Guide*.

<table>
<thead>
<tr>
<th>Class</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>article</strong></td>
<td>This class is by far the most common. It is used for most short documents.</td>
</tr>
<tr>
<td></td>
<td>See Sections 2.2.2 and C.5.1 in <em>The $\LaTeX$ Manual</em>.</td>
</tr>
<tr>
<td><strong>report</strong></td>
<td>This class is generally used for longer documents. See Sections 2.2.2 and</td>
</tr>
<tr>
<td></td>
<td>C.5.1 in <em>The $\LaTeX$ Manual</em>.</td>
</tr>
<tr>
<td><strong>book</strong></td>
<td>This class is meant for actual books. See Sections 5.1 and C.5.1 in <em>The $\LaTeX$ Manual</em>.</td>
</tr>
<tr>
<td><strong>slides</strong></td>
<td>This class can be used in creating originals from which overhead transparencies</td>
</tr>
<tr>
<td></td>
<td>to be used in presentations can be made. The type size, including the size</td>
</tr>
<tr>
<td></td>
<td>used for mathematical symbols, is quite large, and transparencies readable</td>
</tr>
<tr>
<td></td>
<td>from some distance will be produced. See Sections 5.2 and C.5.1 in <em>The $\LaTeX$ Manual</em>.</td>
</tr>
<tr>
<td><strong>letter</strong></td>
<td>This class facilitates creating letters. See Sections 5.3 and C.5.1 in <em>The $\LaTeX$ Manual</em>.</td>
</tr>
</tbody>
</table>

Table A.2: The most common options for the `documentclass` command. Numerous other options are described in Section C.5.1 in *The $\LaTeX$ Manual* and in the other references listed in Section A.18.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>10pt or 11pt or 12pt</td>
<td>(default = 10pt) Specifies default type size.</td>
</tr>
<tr>
<td>letterpaper or legalpaper or a4paper or ...</td>
<td>(default depends on site but is usually letterpaper in the US)</td>
</tr>
<tr>
<td>landscape</td>
<td>(default is portrait orientation) Specifies that text is to be formatted</td>
</tr>
<tr>
<td></td>
<td>for landscape orientation on the specified paper size.</td>
</tr>
<tr>
<td>oneside or twoside</td>
<td>(default = oneside) Specifies whether output is to be formatted for single- or double-sided printing. The twoside option allows for left and right margins and running head positions to be different on odd and even numbered pages.</td>
</tr>
<tr>
<td>onecolumn or twocolumn</td>
<td>(default = onecolumn) Specifies one- or two-column formatting on each page.</td>
</tr>
</tbody>
</table>

While there can be material in the code beyond this final command, none of that material will be read or processed by $\LaTeX$.

The three commands described in the previous paragraph are *mandatory* in all documents. Supplemented by the additional feature that a blank line in the code triggers a new paragraph, they are also *sufficient* for the creation of code for any document consisting of nothing but paragraphs of straight text to be formatted in accordance with all of $\LaTeX$’s defaults. Table A.3 shows a listing of a simple code that invokes only these commands.

While the code is simple (as in Table A.3) or much more elaborate, producing the final printed document involves additional steps. One first creates a printable file—PostScript and PDF are the most common—containing the formatted document and then displays that file on a screen and/or sends it to a printer.
Table A.3: A simple \LaTeX code.

\documentclass{article}
\begin{document}

This sample illustrates the simplest code containing the mandatory commands and a brief text. Since no optional formatting commands have been included, the text will be formatted using all of the built-in defaults (margins, paragraph indent, type size and font, etc.).

The blank line preceding this line in the code will trigger a new paragraph. Each paragraph is indented, but note that there is no extra space between paragraphs. Note also that the lines in the code can be quite ragged; they will be filled and justified in the processing that produces the final copy.

\end{document}

A.1.2 Creating a PostScript File

Converting the \LaTeX source file into a PostScript file describing the desired document involves two steps:

1. First, we must “compile” the code by processing the file with \LaTeX, a task accomplished either (1) by typing a command like

   \texttt{latex filename} \hspace{1cm} (Default extension for the input file is .tex.)

   where \texttt{filename} specifies the input file to be processed, or (2) by selecting an item from a menu in a graphical user interface (GUI).\footnote{See the \textit{Local Guide} for the precise command at your site.} However \LaTeX is invoked, it will create several output files, all of which will have the same name as the input file except for the extension: (1) a binary output file, with extension .dvi, which contains the translation of the original file into \TeX’s \textit{device} independent language; (2) an ASCII log file, with extension .log, which contains an expanded log of the messages that appear on the screen as \texttt{latex} works its magic on the input file; and (3) an auxiliary ASCII file, with extension .aux, which contains information that is important only if the code makes use of \LaTeX’s capabilities for generating internal cross references.\footnote{These capabilities will be described in later sections of this Appendix. For now, we may ignore messages relating to the .aux file, noting only that, when we ultimately \texttt{do} make use of internal references, the \textit{first} pass of the document through \LaTeX writes the .aux file and a \textit{second}—and occasionally a \textit{third}—pass is necessary to incorporate that information in the final formatting of the pages.} The resulting .dvi file can be directly displayed on the screen (see Section A.1.4).\footnote{Remember that we have limited the present discussion to documents containing no figures. See Section A.7 for the adjustments to be made if PostScript or PDF figures are included.}

2. Second, we must translate the .dvi file into a file that can be understood by the printer on which the document is to be printed. Most often, the .dvi file will be translated into
A.1. CREATING A SIMPLE DOCUMENT

Table A.4: The output from the code in Table A.3. The page number at the bottom of the page has been cut off in this display.

This sample illustrates the simplest code containing the mandatory commands and a brief text. Since no optional formatting commands have been included, the text will be formatted using all of the built-in defaults (margins, paragraph indent, type size and font, etc.).

The blank line preceding this line in the code will trigger a new paragraph. Each paragraph is indented, but note that there is no extra space between paragraphs. Note also that the lines in the code can be quite ragged; they will be filled and justified in the processing that produces the final copy.

PostScript, the most common program for effecting that translation being dvips. The necessary translation will be achieved either (1) by typing a command like\footnote{Additional information about dvips can be found by typing the command dvips with no arguments. Further, the option \texttt{-pp} allows selection of specific pages from the full document, the options \texttt{-p} (first page) and \texttt{-l} (last page) allow selection of a range of pages from the full document, and the option \texttt{-t} allows printing of the output in landscape orientation rather than in the default portrait orientation. Even more detailed information is available in the on-line help for dvips, accessed in many systems by typing the statement \texttt{texdoc dvips} at a Shell prompt or, in the UNIX and LINUX operating systems, the statement \texttt{man dvips} at a Shell prompt.}

\begin{verbatim}
dvips -o filename.ps -t letter filename
\end{verbatim}

which translates the entire file into PostScript, stores it in a file, and (with the \texttt{-o} option here used) gives the stored file the same name as the input file but with extension \texttt{.ps} and with the \texttt{-t} option forces use of letter size paper\footnote{If the explicit specification of the output file is omitted, dvips will attempt to send the file directly to the printer—and the whole operation will probably fail. Further, the PostScript file will then not be stored for subsequent use.} or (2) by selecting an item from a menu.\footnote{This option is often the default. The option \texttt{-t landscape} will override the default portrait orientation.}

To be more explicit, the file \texttt{texsample1.tex}, contains the \LaTeX\ code in Table A.3. Once this file has been copied to the default directory, we would produce the \texttt{.dvi} and \texttt{.ps} files with statements like

\begin{verbatim}
latex texsample1
dvips -o texsample1.ps -t letter texsample1
\end{verbatim}

or equivalent selections from a menu.

A.1.3 Creating a PDF File

PDF files can be created from the \LaTeX\ code in at least two ways:

1. Starting with \LaTeX\ source file created in Section A.1.1, one submits to the operating system the single statement\footnote{Check the Local Guide for more details. In particular, some sites may have implemented the shorthand command \texttt{dvip filename} for the command dvips \texttt{-o filename.ps filename}.}

\begin{verbatim}
pdflatex filename
\end{verbatim}

(DEFAULT file type is \texttt{.tex})

or, for the sample above,

\begin{verbatim}
pdflatex texsample1
\end{verbatim}

The program pdflatex is normally installed automatically whenever \LaTeX\ is installed.
which will produce directly a PDF file named filename_.pdf or texsample1.pdf as well as two additional files, specifically the .aux (which we can ignore for now) and the .log file containing an expanded log of the messages that appear on the screen as \texttt{latex} works its magic.

2. Starting with the .ps file produced in Section A.1.2, one executes the single statement

\texttt{ps2pdf filename.ps} \hspace{1cm} (No default file type)

or, for the sample above
\texttt{ps2pdf texsample1.ps}

which will produce a PDF file named filename_.pdf or texsample1.pdf from the .ps file.

Exploration of other routes to a PDF file is left to the reader. In particular, the program \texttt{dvipdfm} can sometimes be used to convert the .dvi file created above directly to a PDF file without creating the PostScript file as an intermediary.

\section*{A.1.4 Displaying the Document on the Screen}

Numerous programs for displaying documents produced with \LaTeX{} on the screen exist. Among the more common are the following:

- (for displaying a .dvi file) \texttt{xdvi} for UNIX and LINUX platforms and \texttt{yap} (\textit{yet another previewer}) for windows platforms. These programs are described more fully in Section A.14.
- (for displaying a PostScript file) \texttt{ghostview}, which is available for almost all platforms.
- (for displaying a PDF file) Adobe \texttt{acroread}, which is available for almost all platforms.

These programs are invoked by double-clicking ML on an icon for the program and then opening the desired file, by double-clicking ML on an icon for the file to be displayed, by executing an appropriate statement as a command in a \textit{Shell} window, or by double-clicking ML on the name of the file in a directory displayed on the screen.\footnote{See the \textit{Local Guide} for specifics at your site.}

\section*{A.1.5 Printing the Document}

Programs (\texttt{xdvi}, \texttt{yap}, \texttt{ghostview}, \texttt{acroread}) for displaying a file on the screen often have an item in the \textit{File} menu to request printing of the file on an available printer, and many times the print utility thereby invoked offers options for double-sided printing, scaling of the output, ... Often, a simple command like

\begin{center}
\begin{tabular}{ll}
UNIX/LINUX/MAC & Windows \\
\texttt{lp filename.ps} & \texttt{print filename.ps} \\
\end{tabular}
\end{center}

submitted from a \textit{Shell} or \textit{Command} window will request a printed copy of the file, though the destination printer must be known to the operating system and able to translate PostScript.\footnote{Printers without this capability are rare.}

In the case of the typed command, the extension must this time be explicitly present, since the programs \texttt{lp} and \texttt{print} make no assumptions about file type.\footnote{Again, check the \textit{Local Guide} for details on how to print a file.}

In any case, the resulting output for our sample file is shown in Table A.4.
A.2 Specification of Global Style: The Preamble

The standard document classes have default settings for the page setup (area allocated to text, margins, paragraph indent, paragraph separation, line spacing, etc.). Only occasionally are these defaults actually appropriate for the document being prepared. Therefore, it is frequently necessary to modify the global style of the document. Commands that accomplish this end are normally placed in the *preamble*—the section between the command `\documentclass` and the command `\begin{document}`—and affect the entire document. The most commonly used commands in the preamble invoke the general command

\[ \text{\texttt{\textbackslash setlength\{LengthParameter\}\{Value\}}} \]


\[ \text{\texttt{\textbackslash setlength\{\textwidth\}\{6truein\}}} \]
\[ \text{\texttt{\textbackslash setlength\{\textwidth\}\{6truein\}}} \]
\[ \text{\texttt{\textbackslash setlength\{\oddsidemargin\}\{0.25truein\}}} \]
\[ \text{\texttt{\textbackslash setlength\{\topmargin\}\{-0.5truein\}}} \]
\[ \text{\texttt{\textbackslash setlength\{\parindent\}\{20pt\}}} \]
\[ \text{\texttt{\textbackslash setlength\{\parskip\}\{6pt plus 2pt minus 1pt\}}} \]

placed in the preamble will change the defaults to specify a 6” × 9” area of text centered on an 8.5” × 11” page with a 20 point paragraph indent and an extra 5–8 points between paragraphs. Indeed, these are (almost\textsuperscript{21}) the settings used for this book.

A few issues of global style are specified by an optional argument in the command `\documentclass`. The most common of these arguments specifies the type size. By default, the document will be set in 10 point type, which is standard for the main text in many journals. To change the size of the characters throughout a document, modify the `\documentclass` command to\textsuperscript{22}

\[ \text{\texttt{\documentclass[TypeSize\{Class\}]\}}} \]

where, in the *article*, *book*, and *report* classes, *TypeSize* can be one of 10pt (the default), 11pt, or 12pt. The *slides* class *admits* these specifications of size, but *ignores* them. The main text in this book is set in \LaTeX{}’s 10 point type, which is also common in many technical journals.

Several other global specifications can be included in the preamble or as optional arguments in the command `\documentclass`, including specifications regarding positioning and style of page headers, definitions of new commands, more subtle changes in the global style, selection of two-column format, specifications to anticipate ultimate printing on both sides of the page, etc. For details, the reader is referred to *The \LaTeX{} Manual*.

Specific instructions for document styles can be saved in files for easy incorporation in documents as they are created. Several possibly useful *templates* are enumerated in Table A.6. Consult the

\textsuperscript{20}The *point*, which is a standard printer’s measure of length, is 1/72” (72 points per inch).

\textsuperscript{21}The difference lies in the specification of the left margins. For this book (which uses double-sided printing), \LaTeX{}’s ability to specify one left margin (with `\oddsidemargin`) for odd-numbered pages and a different left margin (with `\evensidemargin`) for even-numbered pages has been exploited to keep the text more fully out of the binding than would otherwise be the case.

\textsuperscript{22}By convention in \LaTeX{}, *optional* arguments to a command are enclosed in *square brackets* and *mandatory* arguments are enclosed in *curly braces*. 
APPENDIX A. INTRODUCTION TO \LaTeX

Table A.5: Selected parameters that affect the size and positioning of the text area on a page. The given default values apply to the \texttt{article} style. Further information can be found in Section 6.4.1, at the very end of Section C.5.3, and in Fig. C.3 in \textit{The \LaTeX\ Manual}.

\texttt{\texttt{textheight}} (default 7.375") Specifies the vertical dimension of the region of the page occupied by text, excluding the head and the foot of the page. Its value is commonly 9" for an 8.5" × 11" page.

\texttt{\texttt{textwidth}} (default 4.75") Specifies the horizontal dimension of the region of the page occupied by text. Its value is commonly 6" for an 8.5" × 11" page.

\texttt{\texttt{oddsidemargin}} (default 0.75") For all pages (single-sided output) or odd-numbered pages (double-sided output), specifies the left margin—distance from the left edge of the page to the left edge of the region occupied by text—to be 1" plus the specified value. The value 0.25" will center 6" wide text on an 8.5" × 11" page.

\texttt{\texttt{evensidemargin}} (default 0.75") Only for even-numbered pages with double-sided output, specifies the left margin—distance from the left edge of the page to the left edge of the region occupied by text—to be 1" plus the specified value. The value 0.25" will center 6" wide text on an 8.5" × 11" page.

\texttt{\texttt{topmargin}} (default 0.25") Specifies the top margin, i.e., the distance from the top edge of the page to the top edge of the header above the main region occupied by text, to be 1.0" plus the specified value. By default, the height of the header (specified by \texttt{\texttt{headheight}}) and the separation of the header from the main text (specified by \texttt{\texttt{headsep}}) together add to 0.5", so—if the defaults for \texttt{\texttt{headheight}} and \texttt{\texttt{headsep}} are accepted—we can \textit{pretend} that \texttt{\texttt{topmargin}} specifies the distance from the top edge of the page to the top edge of “real” text to be 1.5" plus the specified value. Thus, for example, the value −0.5" will center 9" high text on an 8.5" × 11" page and place the header line—if any—0.5" above the first line of text.

\texttt{\texttt{parindent}} (default 15.0 points) Specifies the paragraph indent.

\texttt{\texttt{parskip}} (default 0.0 points) Specifies the extra space between paragraphs.

\hline

Table A.6: Templates in the directory \texttt{$\texttt{HEAD/tex}$}.

\texttt{lu_article.template} Specifies a general article style, including a 6" × 9" text area centered on an 8.5" × 11" page.

\texttt{lu_cpl.template} Specifies the style of CPL publications.

\texttt{lu_sloan.template} Specifies the two-column style used for the proceedings of the Sloan/Lawrence conference, including capacity for a two-column wide title and abstract on the first page.

\texttt{lu_letter.template} Supplies a starting format for business letters.

\texttt{lu_memo.template} Supplies a starting format for memos.
A.3. In-Text Specification of Local Style

In addition to the global issues discussed in Section A.2, documents are affected locally by many less extensive changes of style. We have already mentioned the use of a blank line to trigger a new paragraph. Commands for accomplishing other common local changes are described briefly in this section.

A.3.1 Type Style

In the terminology of type setters, the phrase *type style* refers collectively to combinations of three independent characteristics of the type. In the most sophisticated description, type style is specified by selecting the *series* (medium, bold), *family* (Roman, sans serif, typewriter), and *shape* (upright, italic, slanted, small cap) of the desired style. Each characteristic is specified independently of the other two. Repeated selection of all three characteristics, however, is cumbersome, so—as enumerated in Table A.7—\LaTeX provides simpler ways to select the more common combinations.

Some of these mechanisms make use of *declarations*, which change the type style until it is explicitly changed again. If, for example, we included the declaration \bf at some point in our code, everything in our document from that point until we change the style with another declaration would be set in bold, Roman, upright type. To limit the scope of this declaration to some portion of the text, the text to be “emboldened”, including the declaration, would be enclosed in curly braces (as shown in the table). Once \LaTeX’s processing has passed beyond the point of the closing curly brace, the declaration is no longer in effect and the type style reverts to what it was before the opening curly brace.

Emphasis of a word or a phrase now and then would, in normal use, be achieved with *italic* type. In \LaTeX, the declaration \it might be used. The declaration \em, however, is preferred because it is sensitive to the current type style. When the current type style is Roman, \em will shift to italic type; when the current type style is italic, \em will shift to Roman type. Thus, \em can be used inside a phrase that is already being emphasized. Emphasized text inside of emphasized text that is itself embedded in Roman text will be set in Roman text to contrast with the italic style of the text that immediately surrounds it.

---

23 The default is medium series, Roman, upright shape—the type style used for this book.
24 Because it is sometimes difficult to remember which “commands” are commands and which are actually declarations, we index both as commands.
Note also in Table A.7 that emphasized text can be achieved not only with the declaration \em but also with the command \textbf. The former form is perhaps preferable for longer phrases; the latter would be used for a word or two. Each achieves the same effect.

To be especially fastidious, we should recognize that return from italic or emphasized type to other type styles may result in too little space between the last italic or emphasized letter and the first Roman (say) letter. The command \textbackslash inserted immediately after the last italicized or emphasized letter instructs \LaTeX to insert the last letter’s italic correction; use of this correction will usually improve the legibility of the final copy. Thus, the careful way to specify an emphasized word is, for example, \texttt{\em and\textbackslash}. (The italic correction is not necessary in all situations, e.g., if the following character is a period or a comma, which explains why it cannot be automatically and always inserted.)

In using one or another of these simple specifications, we are accepting particular combinations of series, family, and shape. \LaTeX provides two ways to take advantage of the full flexibility with separate specification of these three characteristics. For example, we might use declarations to specify bold, sans serif, upright type with the construction

\texttt{\bfseries\sffamily\upshape text}

Alternatively, we might use the commands

\texttt{\textbf{\textsf{\textup{text}}}}

To combine a change in type style with a change in type size, place the specification of type size first, e.g., \texttt{\Large\bf text} to select large bold-face type.
Table A.8: Commands for changing type size. Note that the actual size produced by each command is relative to the global type size (10 pt, 11 pt, or 12 pt) in use. Note also that, in some classes, adjacent members in this sequence may be assigned to the same type size.

<table>
<thead>
<tr>
<th>Command</th>
<th>Function</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>\tiny \text{text}</td>
<td>Sets text in tiny type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\scriptsize \text</td>
<td>Sets text in scriptsize type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\footnotesize \text</td>
<td>Sets text in footnotesize type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\small \text</td>
<td>Sets text in small type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\normalsize \text</td>
<td>Sets text in normalsize type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\large \text</td>
<td>Sets text in large type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\Large \text</td>
<td>Sets text in Large type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\LARGE \text</td>
<td>Sets text in LARGE type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\huge \text</td>
<td>Sets text in huge type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\Huge \text</td>
<td>Sets text in Huge type.</td>
<td>Sample</td>
</tr>
</tbody>
</table>

A.3.3 White Space

Some aspects of the way the components (text, tables, graphs, ...) of a document are placed in the available space are controlled by the length parameters set in the preamble and discussed in Section A.2. Additional features of \LaTeX that give control over this feature of a document include:

- The command \noindent to suppress paragraph indentation for a single paragraph.
- The commands \newpage and \clearpage to force a new page. The first of these commands terminates the current page and starts a new page. The second also flushes out any accumulated tables and figures that have not yet been output.
- The length parameter \baselineskip to specify the linespacing. This parameter is set with the command \setlength described in Section A.2. The default value is 12 points, which yields single spacing. The value 18 points specifies space and a half, and the value 24 points specifies double spacing. This specification must be placed after the command \begin{document}, and it can be changed whenever appropriate to the text. See, however, Section A.3.4 for a comment about limitations of this approach to setting the linespacing.
- The commands \quad and \qquad, which insert an en space and an em space, respectively. (An en space is about the width of the letter x, and an em space is about the width of the letter M, both in the current font.)
- The command \vspace{Value} to generate vertical space. Here, Value can be any defined length parameter, e.g., \vspace{\parskip}, or a specific length, e.g., \vspace{2.0truein}. If the space happens to occur at the top of a new page, it will be omitted, but the command \vspace{Value} will force the space to be included even if it is at the top of a page.
- The command \hspace{Value} to generate horizontal space. Here, Value can be any defined length parameter, e.g., \hspace{\parindent}, or a specific length, e.g., \hspace{36.0pt}. If the space happens to occur at the end of a line, it will be omitted, but the command \hspace{Value} will force the space to be included even if it is at the end of a line.
- The commands \settowidth, \settoheight, and \settodepth, which return lengths for the corresponding dimensions of the box that will enclose whatever is the argument of the
command. No printed output is produced by these commands. We might, for example, define a new length parameter \texttt{dblast} as the width of the box accommodating two asterisks with the two commands

\newlength{dblast}
\settowidth{dblast}{**}

and then insert the space for two asterisks (without actually displaying them) with the command

\hspace{dblast}

These commands are described in detail at the end of Sections 6.4.1 and C.13.1 in \textit{The \LaTeX\ Manual}.

\textbf{A.3.4 Environments}

Sometimes, the formatting to be accomplished requires a more sophisticated definition of a change than is possible within the framework of a simple declaration or command with a few arguments. \LaTeX\'s \textit{environments} are more versatile and flexible than declarations and commands. For example, to present text in a narrower paragraph, as is conventionally done with extended quotations, we might invoke one of the constructions

\begin{quote}
Text of quotation.
\end{quote}

or

\begin{quotation}
Text of quotation.
\end{quotation}

Normally, the command \texttt{\begin{...}} will be placed on a line by itself, the text may spread over several lines, and the command \texttt{\end{...}} will be placed on a line by itself. Left and right margins are indented equally. Unless the text includes its own formatting specifications, however, the first construction will indent each new paragraph and will place \textit{no} extra space between paragraphs while the second will place extra space between paragraphs and will \textit{not} indent each new paragraph.

For a second example, if we wished to center a line or lines, we might invoke the similar construction

\begin{center}
First line \textbackslash Second line \textbackslash Third line \textbackslash ...\end{center}

Normally, the command \texttt{\begin{...}} will be placed on a line by itself, the text may spread over several lines, each line in the final output will in the code be separated from its predecessor with the command \textbackslash, and the command \texttt{\end{...}} will be placed on a line by itself.

These two examples introduce the construction called an \textit{environment}, which in general has the format

\begin{EnvironmentName}
\end{EnvironmentName}

We met the \texttt{document} environment early on. Beyond the \texttt{quote}, \texttt{quote}, and \texttt{center} environments, the most commonly used environments include \texttt{verbatim}, \texttt{flushleft}, \texttt{flushright}, \texttt{itemize}, \texttt{enumerate}, \texttt{list}, \texttt{displaymath}, \texttt{equation}, \texttt{eqnarray}, \texttt{figure}, \texttt{table}, \texttt{tabular}, \texttt{minipage}, \texttt{picture}, and \texttt{thebibliography}. Environments can be thought of as mini-documents within a document. They have default settings, but many of them possess parameters that the user can modify to customize the details of the environment. Further, as described in Sections 3.4 and C.8 in \textit{The
In-Text Specification of Local Style

\LaTeX Manual, individual users can supplement the standard set by defining additional environments suited to the circumstances of the document. Individual exercises in this book, for example, are formatted by a user-defined environment.

The length parameter \texttt{\baselineskip} interacts in interesting ways with the formatting of text contained within environments. Changing \texttt{\baselineskip} will affect only the spacing in the main text of the document, including text in \texttt{\itemize}, \texttt{\enumerate}, and \texttt{\verbatim} environments. Footnotes, material in \texttt{\tabular} environments, table and figure captions, and perhaps other components will remain single spaced unless—as described in Section C.3.2 in \LaTeX Manual—the parameter \texttt{\baselinestretch} is also changed with the command \texttt{\renewcommand}. This additional change, however, has interesting side effects. (For example, the spacing of lines within footnotes will be changed but the spacing between separate footnotes on the same page remains single!) Tampering with global style can at times have unintended (and unwanted) consequences. \textit{Beware!}

\subsection{\LaTeX Packages}

As a publishing system, \LaTeX is infinitely extendable. Rather than incorporate all manner of special tools within the basic program, its designers provided a means by which additional features could be added through the use of packages, the features of which can be made available in any specific code by placing the command

\begin{verbatim}
\usepackage{ PackageName }
\end{verbatim}

in the preamble The \LaTeX Companion describes numerous packages, including \texttt{amstex}, \texttt{babel}, \texttt{color}, \texttt{graphics}, \texttt{graphicx}, \texttt{graphpap}, \texttt{ifthen}, \texttt{latexsym}, \texttt{imakeidx}, \texttt{verbatim}, \texttt{pict2e}, and \texttt{showidx}. A brief description of many of these packages is included in Section C.5.2 of The \LaTeX Manual. The statement texdoc PackageName at a Shell prompt also will bring up documentation on several of these packages.

In addition, some programs, especially those with notebook capabilities, have the ability to write the contents of the notebooks into \LaTeX source files. Usually, those programs make use of their own special \LaTeX commands and supply program-specific \LaTeX packages which must be accessible, either at the proper point in the \LaTeX directory structure or in the directory from which you run \LaTeX when processing a \LaTeX file produced by the programs. Details on those additional packages will be found in the manuals provided by the vendors of the programs.

\subsection{Miscellaneous Other Capabilities}

Several additional capabilities merit particular comment:

\begin{itemize}
  \item To force \LaTeX to keep words together on a single line and/or to prevent \LaTeX from inserting additional space as it justifies a line, use a tilde \texttt{\textasciitilde} instead of a space between the words. The tilde should also be used after a period that does not end a sentence. For example, we should type \texttt{\textasciitilde Mr. \textasciitilde Cook} rather than\texttt{\textasciitilde Mr. \textasciitilde Cook} in our file so as to produce ‘Mr. Cook’ rather than ‘Mr. Cook’ in our output. In the second form, there is more space between the abbreviation and the name because \LaTeX automatically inserts extra space after each period (which is assumed to mark the end of a sentence). \LaTeX may add even more space in the second case as it justifies the line between the margins.
  \item To generate a dash, use \texttt{-}, \texttt{--}, or \texttt{---}, depending on the length of the required dash. One hyphen produces a hyphen, as in two-column; two hyphens in a row produce a slightly longer
\end{itemize}

\footnote{The symbol ‘\texttt{\textunderscore}’ underscores the presence of a space at the indicated point; it is not a character explicitly present in the code.}
dash (an en dash in printers’ terminology), conventionally used to indicate ranges, e.g., 10–12; three hyphens in a row produce a still longer line (—, an em dash in printers’ terminology), sometimes called a punctuation dash and conventionally used in pairs as an alternative to parentheses. Note that none of these constructions is a minus sign −, which has yet a different length and is produced only in math mode. (See Section A.4.)

- To generate opening and closing (double) quotation marks, use ‘‘ and ’’, i.e., two opening or closing (single) tics in a row, respectively. Do not use the symbol * for either of these punctuation marks.

- To specify a footnote, use the command \footnote{Text of footnote.}. The specified text will be placed at the bottom of the page and keyed to the text with an automatically generated number that starts with 1 at the beginning of the document.

- To start a new section, use the command \section{Title of Section}. In the \texttt{article} class, section numbers start with 1 at the beginning of the document and are automatically generated and incremented; the section number is included in the printed section title, which is displayed in a type size and font specified in the document class, and indentation will be suppressed in the first paragraph of each new section. [The commands \subsection and \subsubsection function in a similar way for subsections and subsubsections in the document. In the \texttt{book} class, the commands \chapter and \part (an aggregation of chapters) are also available.]

- To control hyphenation, insert the command \- at the point in a word where \LaTeX{} is permitted to insert a hyphen as it fills and justifies lines. \LaTeX{} has its own rules for hyphenation, but they are not infallible. Sometimes \LaTeX{} needs help. Note that words containing even one indication of an optional hyphen will not be hyphenated at any other point(s) in the word.

- To suppress hyphenation of a single word altogether, place the word as the argument of an \texttt{mbox} command, e.g., \texttt{mbox{customize}}.

### A.4 Including Equations

Most scientific documents will have equations, all of which must be specified in \LaTeX{}’s math mode. Shifting from the default text mode to math mode can be accomplished in several ways. For short, unnumbered equations or mathematical symbols that appear in a line of text, the equation or symbol must be enclosed in one of the two constructions \(\ldots\) or \$.\ldots\$.\$, which are equivalent. If we want one or more displayed equations, with or without automatically generated equation numbers, we need to use the form \texttt{\begin...\end}. Three standard environments use this construction. The \texttt{displaymath} environment centers a single equation on a line by itself, does not number the equation, and can be specified by the shortened form \texttt{\[\ldots\]} . The \texttt{equation} environment centers a single equation on a line by itself and numbers the equation automatically. The \texttt{eqnarray} environment—see \textit{The \LaTeX{} Manual}—is used for a string of equations and for long equations that will not fit on one line; its syntax allows vertical alignment of two or more displayed and numbered equations. To facilitate references to numbered equations produced in the \texttt{equation} and \texttt{eqnarray} environments, \LaTeX{} provides the command \texttt{\label{ReferenceName}} for defining a symbolic label within the environment and the command \texttt{\ref{ReferenceName}} to permit referencing the equation by its number in the text. Further, the command \texttt{\pageref{ReferenceName}} permits referencing the page on which the equation occurs in the text.\footnote{26}{Warning: The \texttt{displaymath}, \texttt{equation}, and \texttt{eqnarray} environments will generate error messages if they contain blank lines.}

\footnote{When these references are used, the information written into the .aux file becomes important and the code must be processed by \texttt{latex} or \texttt{pdflatex} twice, once to write the correct information into that file and a second time to read and make use of that information. Other environments, such as \texttt{table} and \texttt{figure} also admit a \texttt{\label} command and can make use of this capability for symbolic internal references to those components. The label command can also be used inside the argument of \texttt{\part}, \texttt{\chapter}, \texttt{\section}, \texttt{\subsection}, \texttt{\subsubsection}, and \texttt{\footnote} and in the text following an \texttt{\item} command in an \texttt{enumerate} environment to facilitate reference to these components of the document.}
A.5. INCLUDING LISTS

\LaTeX also has many symbols and structures, common in mathematical formulas, that can be used only in math mode. Subscripts and superscripts, roots, Greek letters, integral signs, summation signs, the \texttt{array} environment, and many more are available.\textsuperscript{27} To reiterate, these symbols and structures can only be used if \LaTeX is in math mode. So, to insert the symbol $\Theta$ in the final output, we place the command $\$\Theta\$ in the code.\textsuperscript{28} (Here, the first dollar sign puts \LaTeX in undisplayed math mode and the second returns \LaTeX to text mode.) Since \LaTeX follows its own rules about spacing when in math mode, spaces in math mode are usually ignored. (A space, however, is necessary in such contexts as $\omega t$ to separate the command $\omega$ from the character $t$.) The commands $\$, $\!$, $\,$, and $\;$ can be used in math mode to fine tune the spaces between elements in the equation by adding a thin space, a negative thin space, a medium space, and a thick space, respectively. It is, for example, customary to separate a differential element from what precedes and follows with a thin space, e.g., $(x + y)\, dx\, dy$ rather than $(x + y)\, dx\, dy$.\textsuperscript{29}

Following the conventions for the type setting of mathematical text, all symbols in a math environment are automatically italicized. The most common departure from this convention occurs with the names of the standard mathematical functions, which are conventionally set in Roman type. Special math mode commands of the form $\cos$ and $\sin$ should be used for these functions.\textsuperscript{30} Except for $\sc$, the declarations shown in Table A.7 can also be used in math mode to specify text in something other than italic type but, in some contexts it may be preferable to use one of the commands $\mathrm$, $\mathit$, $\mathbf$, $\mathtt$, etc. described in Section 3.3.8 of The \LaTeX Manual. Remember, however, that spaces are ignored in math mode, so a shift to Roman type with the command $\mathrm$, for example, to include a two-word phrase, will result in the words running together unless the space command $\,\,$ is used to insert an explicit space between the words. Sometimes, the better means to insert Roman text in equations is to use the command $\mbox$ to escape temporarily to text mode.

Including bold characters in equations is especially tricky. The command $\mathbf$ “emboldens” only \emph{some} characters in the processed formula and, in particular, leaves lower case Greek symbols in light face. The $\boldmath$ declaration causes \emph{everything} in math mode to be bold but cannot itself be \emph{invoked} in math mode. To include a bold face, lower case $\theta$, for example, we must use the construction

\[
\$ \ldots \mbox{ \boldmath \$\theta\$ } \ldots \$
\]

where the command $\mbox$, which \emph{is} valid in math mode, temporarily shifts to text mode so the $\boldmath$ declaration can be invoked and the math mode expression in the argument of the command $\mbox$ will then be made bold; the bold presentation is confined by the command $\mbox$ to the desired part of the total expression.

A.5 Including Lists

Among the most commonly used environments are those that facilitate creation of lists, including the \texttt{itemize} environment for bulleted lists (such as the one in Section A.3.6) and the \texttt{enumerate} environment for lists in which the items are lettered or numbered in sequence (such as the one in Section A.17). Each environment is introduced and terminated with the standard construction $\texttt{\begin} \ldots \texttt{\end}$. Within the environment, each new item is introduced with the command $\texttt{item}$. Thus, the structure might look like

\textsuperscript{27}For a complete list of all symbols and the commands that create them, see Section 3.3 on Mathematical Formulas in The \LaTeX Manual.

\textsuperscript{28}In math mode, Greek letters can be produced by commands that simply name the letter, e.g., $\Theta$ for $\Theta$ and $\omega$ for $\omega$. Only the initial letter of the name is capitalized to produce an upper-case letter. Note, however, that commands for letters that are identical to Arabic letters, e.g., $\kappa$, do not exist.

\textsuperscript{29}Technically and officially, the differential element $dx$ should be written with a \textit{Roman} d and an \textit{italic} x, i.e., $dx$, but this convention is rarely followed.

\textsuperscript{30}See Table 3.9 in The \LaTeX Manual for a complete listing.
\begin{itemize}
\item Text of first item.
\item Text of second item.
\item Text of third item.
\end{itemize}

The bullets and the numbers are, of course, generated automatically and, within the `enumerate` environment, the numbers are automatically incremented. Details will be found in Sections 2.2.4, 6.6, C.6.2, and C.6.3 in *The \LaTeX{} Manual*. Note in particular

- The optional argument for the command \texttt{\item}, which provides a means to override the automatic label for that item and replace it with an explicit stipulated label. For example, the command \texttt{\item[DMC]} will cause the item to be labeled ‘DMC’ and the command \texttt{\item[]} will suppress the label altogether. Note that, if the text of the item itself begins with something enclosed in square brackets, the command introducing that item must be written \texttt{\item{}} to prevent the text in square brackets from being interpreted as the desired label.

- The length parameters \texttt{\itemsep}, \texttt{\parskip}, and \texttt{\parsep}, which provide a means to override default spacings within the environment.

- The way to change the “bullets” in the `itemize` environment. The symbols used to label the “bulleted” items in the various levels of nested `itemize` environments are created by the commands \texttt{\labelitemi}, \texttt{\labelitemii}, \texttt{\labelitemiii}, and \texttt{\labelitemiv}. The default sequence for marking items at each level is •, – *, and ·. Each of these symbols can, however, be changed by redefining the corresponding command. For example, the command

\begin{verbatim}
\renewcommand{\labelitemii}{\circ}
\end{verbatim}

executed in the preamble will change the symbol – to ◦ for the second level in nested `itemize` environments.

- The way to change the form of the labels for each item in the `enumerate` environment. The labels at the various levels of nested `enumerate` environments are determined by invoking one of the commands \texttt{\theenumi}, \texttt{\theenumii}, \texttt{\theenumiii}, and \texttt{\theenumiv}, whose action is in turn determined from the value of the corresponding one of the counters \texttt{enumi}, \texttt{enumii}, \texttt{enumiii}, and \texttt{enumiv}. The default sequence for marking items at each level is ‘1.’, ‘(a)’ ‘i,’ ‘A.’, i.e., arabic number with period, lower-case letter in parentheses, lower-case Roman number with period, and upper-case letter with period. These defaults, however, can be changed with a command like

\begin{verbatim}
\renewcommand{\theenumi}{(\Alph{enumi})}
\end{verbatim}

which will change the labels used at the first level in `enumerate` environments to upper-case arabic letters in parentheses. The command \texttt{\Alph} could be replaced with \texttt{\alph}, \texttt{\roman}, \texttt{\Roman}, or \texttt{\arabic} to translate the underlying counter into lower-case arabic letters, lower-case Roman numbers, upper-case Roman numbers, or arabic numbers, respectively. As it turns out, whatever is specified in the argument of the commands translating the counters, a period will be appended by \LaTeX{} at the first, third, and fourth levels, and enclosing parentheses will be supplied at the second level.

### A.6 Including Tables

Creation of columnar arrangements is facilitated by \LaTeX{}’s `tabular` environment, the details of which are involved and are fully described in *The \LaTeX{} Manual*. The `tabular` environment can
A.6. INCLUDING TABLES

Table A.9: Format for inserting a table.

\begin{table}
\caption{...}
\label{...}
\begin{center}
\begin{tabular}{clr}
Row 1, Col 1 & Row 1, Col 2 & Row 1, Col 3 \\
Row 2, Col 1 & Row 2, Col 2 & Row 2, Col 3 \\
Row 3, Col 1 & Row 3, Col 2 & Row 3, Col 3
\end{tabular}
\end{center}
\end{table}

be invoked anywhere in the code and the resulting table will be placed at the point at which the environment appears, perhaps even in the middle of a line of text. Conventionally, however, tables are placed at the top or bottom of the page,\textsuperscript{31} and \LaTeX{} provides the \texttt{table} environment to facilitate proper placement of a table,\textsuperscript{32} though the bare \texttt{table} environment contains nothing other than its name that suggests its use for tables. Any text at all can be incorporated in the \texttt{table} environment and will be placed on the page as would a table. Most commonly, however, the \texttt{table} environment will embrace a \texttt{tabular} environment defining the table itself and will use the command \texttt{\caption} to specify the caption of the table and the command \texttt{\label} to specify a symbolic label for use in referring to the table within the document.\textsuperscript{33} Frequently, the \texttt{tabular} environment will itself be embraced in a \texttt{center} environment to control the horizontal placement of the table on the page. Thus, the common structure for the code defining a table would be of the form shown in Table A.9.\textsuperscript{34,35}

Here, individual elements in a row are separated from one another by ampersands, and the end of each row—except the last—is marked with the command \\.

The illustrative argument \texttt{clr} of the \texttt{tabular} environment requires a bit more explanation.

\begin{itemize}
\item[\textsuperscript{31}] Actually, the command \texttt{\begin{table}} has an optional argument (\texttt{\begin{table}[OptArg]}, which can assume any of the values \texttt{h}, \texttt{b}, or \texttt{t} for placement of the table at the place where the \texttt{table} environment appears (\texttt{h}, for here) or at the bottom (\texttt{b}) or top (\texttt{t}) of the page. In the absence of an explicit specification, \LaTeX{} makes its own decision about placement.
\item[\textsuperscript{32}] If there is space on the current page when the \texttt{table} environment is encountered, the table will be placed on that page. Otherwise, the table will be placed on a subsequent page.
\item[\textsuperscript{33}] The \texttt{\caption} and \texttt{\label} commands can be placed anywhere within the \texttt{table} environment. Captions will commonly be placed either above or below the captioned component. In some contexts (see the warning at the end of Section A.10), placing the caption above the item is preferable.
\item[\textsuperscript{34}] \LaTeX{} automatically uses the specified caption as an entry for a list of tables. There is, however, a limit to the length of such entries. Especially long captions may generate error messages. To avoid this problem, we can—and sometimes must—exploit an optional argument to the command \texttt{\caption{...}} which allows the user to dictate the entry made to the list of tables. Unless a list of tables is to be generated, some authors argue that we should routinely specify a null value for the optional argument by writing the command \texttt{\caption{}{...}}, though we shall here not follow that recommendation. See \textit{The \LaTeX{} Manual} for details.
\item[\textsuperscript{35}] See Sections 3.6 and C.10 in \textit{The \LaTeX{} Manual for much} more detail on ways to line things up in columns. Note particularly (1) the command \texttt{\multicolumn}, which provides for entries that span more than one column, and (2) ways to create horizontal and vertical rulings in the table.
\end{itemize}
In general, this argument will be a string of c’s, l’s and r’s, one for each column in the table. Each specifies the position (centered, left justified, right justified), respectively, of the entry in the corresponding column. In the example, there are three columns, with entries in the first column centered, entries in the second column left justified, and entries in the third column right justified. The argument is mandatory but will, of course, be a string appropriate to the table being constructed rather than the specific string clr.

By default, tables will be produced with no vertical or horizontal rulings. See Section C.10.2 in The \LaTeX\ Manual for a description of the means to add these rulings.

\section{Including Illustrations}

Graphics displays frequently appear in technical documents. Within \LaTeX, means to incorporate graphics include\footnote{We elect here to discuss only a very few of the numerous available packages, choosing those of the broadest applicability. Packages for drawing Feynman diagrams, organic molecules, musical scores, and many other displays are described in The \LaTeX\ Graphics Companion.}

- A cut and paste method (Section A.7.1).
- A method for incorporating a properly constructed PostScript or PDF file that invokes the command $\texttt{\textbackslash includegraphics}$ from the package graphicx (Section A.7.2).
- A method that exploits the \texttt{tikz} package of macros that facilitate describing the graphical display and its formatting in the same way that \LaTeX\ itself facilitates describing the text and its formatting (Section A.7.3).
- The built-in \texttt{picture} environment (Section A.7.4), which has limited capability but is sometimes just the ticket for simple displays.

Most often, commands incorporating figures will be bracketed in a \texttt{figure} environment, which facilitates locating the illustration at the top or bottom of the current (or a following) page.\footnote{The \texttt{figure} environment also admits the optional argument described in footnote 31.} In most cases, we need to know the final vertical extent of the figure in order to specify the size of an appropriate space. Further, the commands $\texttt{\caption}$ and $\texttt{\label}$ are available to caption the figure and to place appropriate entries in the .aux file so that the commands $\texttt{\ref}$ and $\texttt{\pageref}$ will function here as described in the first paragraph of Section A.4.

\subsection{Using Cut and Paste}

The cut and paste method requires only that \LaTeX\ be instructed to set aside appropriately sized and positioned space into which the illustration will be placed after the document has been printed. The segment of the code that creates and labels space for a figure will have the general form

\begin{verbatim}
\begin{figure}
  \caption{ ... }
  \label{ ... }
  \vspace{ ???truein }
\end{figure}
\end{verbatim}

As illustrated here, the caption, including an automatically generated phrase and number of the form ‘Figure 3:’, will appear above the space left for the figure. Positioning the commands $\texttt{\caption}$ and $\texttt{\label}$ after the command $\texttt{\vspace}$ would place the caption below the figure. Note also that,
whichever position is adopted, the command \texttt{label} must appear 	extit{after} the command \texttt{caption}. For this method, you need only have printed copies of your figures.\footnote{Appropriately translated, the substance of footnote 34 applies here to the entry placed automatically in the list of figures.} The format of the files used for the figures is irrelevant and the document can be produced with any of the methods described in Sections A.1.2 and A.1.3.

### A.7.2 Using the graphicx Package

Incorporation of graphic images in a \LaTeX{} document is viewed by \LaTeX{} as the responsibility of the device driver. To enable that process, \LaTeX{} allows the passing of commands to the device driver. Both the format of such commands and the variety of capabilities thus available depends entirely on the device driver. In essence, this process uses the \TeX{} (not \LaTeX{}) command \texttt{special}, but the user is unaware of that underlying command because it is invoked behind the scenes.

The simplest means for importing graphical displays into a \LaTeX{} document—i.e., of commanding \LaTeX{} to pass the proper information on to the device driver—exploits the \texttt{graphicx} package,\footnote{The \texttt{graphics} package might also be used, but the syntax of the commands it defines is marginally less convenient than the syntax of the parallel commands in the \texttt{graphicx} package, i.e. the \texttt{extended graphics} package.} whose features are made available by including that package with the statement

\begin{verbatim}
\usepackage{graphicx}
\end{verbatim}

in the preamble of our \LaTeX{} code.\footnote{More generally, this statement admits an optional argument to the desired graphics driver explicitly. Frequently, that driver will be \texttt{dvips} and the statement would be \texttt{usepackage[dvips]{graphicx}}, but numerous other drivers exist, and many may be installed at your site. See the \textit{Local Guide} for more information. For ultimate flexibility in the formatting of graphics files, we elect to leave that choice to the default.}

After placing the command \texttt{usepackage{graphicx}} in our preamble, we incorporate the graphical display itself by placing a statement like

\begin{verbatim}
\includegraphics[height=\texttt{dimension}]{\texttt{filename}}
\end{verbatim}

at the point where the PostScript or PDF figure defined in by the file \texttt{filename} is to be inserted.\footnote{If you use \texttt{latex}, \texttt{dvips}, and \texttt{ps2pdf} as described in Section A.1.2 and item 2 in Section A.1.3 to produce PostScript and PDF documents, your figures must be available as .ps or .eps files, and—even if the file type is omitted in the \texttt{includegraphics} command—the processing will correctly identify and incorporate the figures. If you seek to produce a PDF document directly with \texttt{pdflatex} as described in item 1 in Section A.1.3, then—even if the file type .pdf is omitted in the \texttt{includegraphics} command—the processing will correctly identify and incorporate the figures. If your figures are PostScript, you must use the first route to the final document; if your figures are PDF, you must use the second route. Software to convert PostScript figures to PDF figures and \textit{vice versa} is described in Section A.11. If both PostScript and PDF files exist for all figures \textit{and} you omit the file type in the \texttt{includegraphics} command, then you can use either \texttt{latex} or \texttt{pdflatex} for the processing.} In general, the image would be centered horizontally on the page by inserting this command in a \texttt{center} environment, e.g.,

\begin{verbatim}
\begin{center}
\includegraphics[height=\texttt{dimension}]{\texttt{filename}}
\end{center}
\end{verbatim}

For example, the code\footnote{Appropriately translated, the substance of footnote 34 applies here to the entry in the list of figures.}
Figure A.1: Mesh surface representation of the irradiance produced by a square aperture.

\begin{figure}
\caption{Mesh surface representation of the irradiance produced by a square aperture.}
\label{LATEX:irrad}
\begin{center}
\includegraphics[height=3.0truein]{diffract}
\end{center}
\end{figure}

will produce Fig. A.1 if the file \texttt{diffract} is stored in the current default directory.\footnote{Omission of the file type in the statement including the graph prepares the way to use either \texttt{latex-dvips} to produce a PostScript file using \texttt{latex} and \texttt{dvips} (Section A.1.2), in which case the file \texttt{diffract.ps} or \texttt{diffract.eps} will be used, or \texttt{pdflatex} to produce a PDF file (Section A.1.3), in which case the file \texttt{diffract.pdf} will be used. The appropriate file must, of course, be accessible in either case.} Here, the specification of the parameter \texttt{height} in the optional argument of the command \texttt{\includegraphics} causes the figure to be scaled to the specified dimension vertically and scaled by the same fraction horizontally to preserve its aspect ratio. As in the example in Section A.7.1, the figure here has been inserted \textit{before} the command \texttt{\caption}, so that the caption will appear \textit{above} the figure. Further, the command \texttt{\label}, which must \textit{follow} the command \texttt{\caption}, defines a symbolic label that can be used within the document to refer to the figure.

\textbf{Warning for OCTAVE users:} All graphics toolkits in OCTAVE will produce proper on-screen graphs and will output PostScript files of these graphs that can be displayed with ghostview. If, however, the graph is to be incorporated in a \TeX document by the procedures described above, graphs produced by all toolkits will be properly incorporated in the \texttt{.dvi} file but those produced by the \texttt{qt} toolkit \textit{may} not then translate with \texttt{dvips} to the subsequent PostScript file. When the latter is the objective, use \texttt{gnuplot} (or maybe \texttt{fltk}). A bit of testing may be necessary.

The command \texttt{\includegraphics} admits several optional arguments. The argument \texttt{width} can also be included. If \texttt{width} is specified \textit{instead of height}, the height will be scaled to preserve

\footnote{The label following the word ‘Figure’ in the output will, of course, reflect the document class in use and the unit of the document in which the figure appears.}
the aspect ratio. If both `height` and `width` are specified, each will be respected and the aspect ratio of the figure may be distorted; specification of `neither` will cause the display to be produced in its original size. Arguments that allow overriding of the bounding box specified in the PostScript file and arguments that permit clipping a portion of a fuller figure are described in *The \LaTeX Graphics Companion*.

### A.7.3 Using the \texttt{tikz} Package

The `tikz` package, which is routinely included in present-day \LaTeX distributions, provides an assortment of \LaTeX macros to facilitate the drawing of simple—and even complicated—figures. Basically, the statement `\usepackage{tikz}` in the preamble followed in the document itself by statements embedded in a `tikzpicture` environment will specify the desired graph and incorporate the graph in a PDF or Postscript file created as described in Sections A.1.2 and A.1.3.\footnote{Note that, when displayed on the screen with \texttt{xdvil} or \texttt{yap}—see Section A.1.4—the intermediate `.dvi` file produced in Section A.1.2 will not contain the correct figures.} Be aware that the code inserted in the `tikzpicture` environment describes the picture in the same way that \LaTeX provides the text and its formatting. Just as one imagines the final formatted document as the \LaTeX coding is constructed, so one has to imagine the resulting picture as one constructs the descriptive code to create it in the final document. \texttt{Tikz} is \textit{not} a WYSIWIG drawing program like \texttt{PAINT} or \texttt{TGIF}. One of its advantages over the route described in Section A.7.2, which imports figures produced in other programs, is that the route of this section guarantees that the fonts used in the figures will be identical to those used in the rest of the document.

The graphics system provided by \texttt{tikz} is impressively elaborate and versatile. As such, learning its capabilities, especially its more sophisticated capabilities, will require substantial effort. Typing the command\footnote{The designation \texttt{pgf}—Portable Graphics Format—reflects the name of the engine underlying the entire system.} \texttt{texdoc pgf}

at a \textit{Shell} window to your operating system will probably bring up links to a number of manuals, including the main—and voluminous (1100-plus pages!)—manual in the file `pgfmanual.pdf`. Further, googling `tikz` will bring up links to numerous documents, including in particular a link to `pgfmanual.pdf` and a link to the much more compact tutorial-style introduction in the file `minimaltikz.pdf`. Once you have selected any of these items, you may in some operating systems have to look in your \textit{Downloads} folder to access the item.

This section quickly orients you to the general features of \texttt{tikz} without in any way pretending to be complete. To that end, having included the statement\footnote{Some installations may also require inclusion of the packages \texttt{pgf} and \texttt{xcolor}.} \texttt{\usepackage{tikz}}

in the preamble of a document, we produce Fig. A.2 by placing the coding in Table A.10 at the appropriate point in the document itself. This coding briefly illustrates some of the more elementary capabilities of \texttt{tikz}. Note specifically the following:

- Optional arguments to any \texttt{tikz} command are included in (square) brackets.

- By default, specified coordinates are expressed in centimeters, though any recognized \LaTeX unit of length can be specifically stipulated. The optional argument `[x=1.0in,y=1.0in]` following `\begin{tikzpicture}` changes the default to inches. Once a figure has been defined, its size can be easily altered simply by changing this optional argument.
• Every statement must end with a semicolon. Any statement can be spread over several lines with only the last line so terminated.

• The statement \textbackslash draw [fill=black] (0,0) circle [radius=0.1]; draws a black-filled circle centered at the origin and having radius 0.1 inches, thus marking the origin of the coordinate system.\footnote{Note that the default unit—here inches—for radius can be overridden with an explicit stipulation, e.g., [radius=0.1cm].} Available colors include red, green, blue, cyan, magenta, yellow, and many others. You can also define your own color using a command like\footnote{The command \textbackslash definecolor here used is defined within the \texttt{tikzpicture} environment and does not require use of the \LaTeX{} package \texttt{color}.}

\textbackslash definecolor\{ColorName\}\{rgb\}\{r,g,b\}

where \textit{ColorName} is the name you assign to the color and \textit{r}, \textit{g}, \textit{b} specify the \textit{r}, \textit{g}, and \textit{b} components that make up the color. Available shapes include \texttt{rectangle}, \texttt{ellipse}, and \texttt{arc}.\footnote{See the manuals for the details of how to specify the dimensions of these shapes.}

• The statement \textbackslash draw [\textbf{<->}, line width=2] (0,1.2)--(0,0)--(2.3,0); draws a line connecting the specified points in the order given, i.e., draws the axes intersecting at the origin. Any number of points can be included in the path. The optional argument \textbf{<->} places an arrow at both ends of the line,\footnote{The separate arguments \textbf{-} and \textbf{->} will place an arrow at the end or the beginning of the line, respectively.} and the argument \texttt{line width} species the weight of the line, by default, in points.\footnote{The argument \texttt{line width=2} could be replaced, among others, by \texttt{ultra thin}, \texttt{very thin}, \texttt{thin}, \texttt{thick}, \texttt{very thick}, and \texttt{ultra thick}, and the default unit could be overridden with an argument like \texttt{line width=0.1cm}.}

• Text is placed where specified with the statements \textbackslash node at (0,1.3) \{\Large\textbf{y}\}; and \textbackslash node at (2.4,0.0) \{\Large\textbf{x}\};. Note that \LaTeX{} stipulations of font size and style are recognized by \texttt{tikz}. Note also that the text is, by default, \textit{centered} at the specified point. See the \texttt{TIKZ} manuals for ways to override that default positioning.

• The statement \textbackslash draw [ultra thick, domain=0:2.0] plot (\textit{x}, \{\textsf{sin(pi*\textit{x} r)}\}) ; illustrates how to draw a smooth curve defined by a function, many of which are available within \texttt{tikz}. Note also the availability of the irrational number \pi{} with the simple coding \texttt{pi}.\footnote{Another way to draw smooth curves is to use a calculational aid in another program to generate a probably long list of coordinates for many points on the curve and then to insert that list as the argument of a \textbackslash draw command.}

• The construction

\begin{verbatim}
\foreach \textit{x} in {0.0,1.0,2.0} {
    \textbackslash draw (\textit{x},0.1 )--(\textit{x},-0.1);
    \textbackslash node at (\textit{x}, -0.4){\textit{x}};};
\end{verbatim}

creates a loop executing the statements enclosed in \{\ldots\} for each value in the list following the keyword \texttt{in}. The two loops in this code place tick marks on the axes and label those marks. Note the semicolons in this construction.

Clearly the command \textbackslash draw is a versatile command that admits numerous embellishments through the use of keywords and/or optional arguments. Conveniently, \texttt{tikz} determines the size of the space to be used in the document to reflect the size of the picture as constructed with the measures provided in the \texttt{tikz} coding.

When features of the package \texttt{tikz} are invoked in the \LaTeX{} code, producing the formatted document must be done carefully. If there are no complications (external figures, table of contents, internal references) to be incorporated, the simple statement
Figure A.2: An illustrative \texttt{tikz} figure. Note that specified colors will be converted to a gray scale unless the display device—screen or printer—can display colors.

```
\begin{tikzpicture} [x=1.0in,y=1.0in]
\draw [fill=black] (0,0) circle [radius=0.1];
\draw [->, line width=2] (0,1.2)--(0,0)--(2.3,0);
\node at (0,1.3) {y};
\node at (2.4,0.0) {x};
\draw [ultra thick, domain=0:2.0] plot (\x, {sin(pi*\x r)});
\foreach \x in {0.0,1.0,2.0} {
  \draw (\x,0.1)--(\x,-0.1);
  \node at (\x, -0.4){\x};}
\foreach \y in {0.0,0.5,1.0} {
  \draw (-0.1,\y)--(0.1,\y);
  \node at (-0.3, \y) {\y};
}\end{tikzpicture}
```

Table A.10: Coding to produce Fig. A.2.

\begin{verbatim}
\begin{figure}
\caption{An illustrative \texttt{tikz} figure.}
\label{LATEX:tigzfig}
\begin{center}
\begin{tikzpicture}
\draw [fill=black] (0,0) circle [radius=0.1];
\draw [->, line width=2] (0,1.2)--(0,0)--(2.3,0);
\node at (0,1.3) {y};
\node at (2.4,0.0) {x};
\draw [ultra thick, domain=0:2.0] plot (\x, {sin(pi*\x r)});
\foreach \x in {0.0,1.0,2.0} {
  \draw (\x,0.1)--(\x,-0.1);
  \node at (\x, -0.4){\x};}
\foreach \y in {0.0,0.5,1.0} {
  \draw (-0.1,\y)--(0.1,\y);
  \node at (-0.3, \y) {\y};
}\end{tikzpicture}
\end{center}
\end{figure}
\end{verbatim}

\texttt{pdflatex filename} \hspace{1cm} (Default file type \texttt{.tex})

will produce a PDF file and the statements

\texttt{latex filename} \hspace{1cm} (Default file type \texttt{.tex})
\texttt{dvips -o filename.ps -t letter filename} \hspace{1cm} (Default file type \texttt{.dvi})
\texttt{ps2pdf filename.ps}

will produce a PostScript and then a PDF file, though figures defined by \texttt{tikz} will not be properly
rendered in the intermediate .dvi file. If a table of contents or internal references or both are involved, two—and maybe three—passes through \texttt{latex} or \texttt{pdflatex} will be necessary. If there are PostScript figures and \textit{no} PDF figures to be incorporated, only \texttt{latex} will work; if there are PDF figures (and \textit{no} PostScript figures) or hyperlinks to be incorporated, only \texttt{pdflatex} will work. Finally, if there are both PostScript and PDF files defining figures, files of one type will have to be converted to the other type before either of these routes to a finished document will work.\footnote{See Section \ref{sec:latexclass} for details about that conversion.}

If you wish to short circuit learning detailed \texttt{tikz} code, you might wish to explore a WYSIWYG editor \texttt{TikzEdt} that provides a graphical interface for creating figures and translates that figure into the corresponding \texttt{tikz} code. This program, which is free, can be downloaded from the web site \url{www.tikzedt.org}. Documentation is also available at that site.

This section provides a woefully incomplete introduction intended more to wet your appetite than to make you an expert. The effort invested to study the manuals identified in the second paragraph of this section will be richly rewarded.

\subsection*{A.7.4 Using the picture Environment and \texttt{pict2e} Package}

Finally, we merely mention the \texttt{picture} environment, described in Sections 7.1 and C.14.1 of \textit{The \LaTeX{} Manual}, and the supplementary \texttt{pict2e} package, described in \textit{The \LaTeX{} Companion}. These components add several commands that facilitate the detailed construction of at least simple graphical displays.

\section*{A.8 Including a Table of Contents, a List of Figures, and a List of Tables}

A table of contents, a list of figures, and a list of tables can be included in the formatted document by the commands \texttt{\tableofcontents}, \texttt{\listoffigures}, and \texttt{\listoftables}. Each

- results in the output of an auxiliary ASCII file of information with file type \texttt{.toc}, \texttt{.lof}, and \texttt{.lot}, respectively,
- requires a second pass through \texttt{latex} or \texttt{pdflatex} to incorporate the list in the finished document, and
- places the list in the output at the point at which the command appears.

When these commands are invoked, each of the components of the text defined by the commands \texttt{\part}, \texttt{\chapter}, \texttt{\section}, \texttt{\subsection}, and \texttt{\subsubsection} will automatically generate a line in the \texttt{.toc} file.\footnote{Note that not all of these options are available in all document classes. For example, \texttt{\part} and \texttt{\chapter} are available only in the \texttt{book} class. Note also the counters \texttt{secnumdepth} and \texttt{tocdepth} that control the depth to which sections are numbered and the depth to which sections are catalogued in the table of contents. The default values of these counters can be overridden using \texttt{\setcounter} in the preamble.} Each use of the \texttt{\caption} command in a \texttt{\figure} or \texttt{\table} environment will automatically generate a line in the \texttt{.lof} or \texttt{.lot} file. Then, in the second pass—which will always be necessary—through \texttt{latex} or \texttt{pdflatex}, these files are read and each line results in an entry in the corresponding list.

Two commands allow explicit entry of information into one or another of these files at the point where the command is inserted in the \LaTeX{} source file, specifically

- the command \texttt{\addcontentsline{file}{unit}{entry}}
- the command \texttt{\addtocontents{file}{text}}
A.9 Including an Index

Here, file is one of toc, lof, and lot, unit is one of part, chapter, section, subsection, or subsubsection if file is toc, figure if file is lof, and table if file is lot.

These two commands have quite different effects. The first adds a bona fide entry to the corresponding file. For example, the line

\addcontentsline{toc}{subsection}{Specially marked point}

will add to the table of contents an entry that is left-justified at the subsection level, contains the text “Specially marked point”, and includes a row of dots and the appropriate page number.\footnote{If you are creating a linked document, this entry in the table of contents will also appear in the navigation panel of that document.} The second merely adds text—no dots or page number—to the corresponding file. For example, the line

\addtocontents{toc}{\vspace{24pt}This is a note.\vspace{24pt}}

inserts a note preceded and followed by a bit of extra vertical space. Any textual note is left-justified as dictated by the section level at which it appears in the document.

Additional details about the issues discussed in this section are laid out in Sections 4.1 and C.4 of The \LaTeX\ Manual. Note, in particular, the optional argument in the command \caption and the several sectioning commands (\chapter, \section, \...), which provide control over the entry each command places in the table of contents, list of figures, and list of tables.

A.9 Including an Index

Among many capabilities, \LaTeX is able to generate an index, though not automatically. In essence, we must place the commands\footnote{Earlier versions of \LaTeX used the package makeidx, which is still available. We here recommend using imakeidx because it makes available a few more features than are available in the previous package and it eliminates the need to run an auxiliary program to format the index.}

\usepackage{imakeidx}
\makeindex

in the preamble and the command

\printindex

at the point in the code at which the index is to be printed—usually the very end. Then, we indicate what items are to be indexed by placing commands like

\index{Item to be indexed}

at appropriate points throughout the document. (The detailed structure of the argument to the command \index is described in Section 4.5 and Appendix A of The \LaTeX\ Manual.) When the file filename.tex is processed by latex or pdflatex, the presence of these commands results in the addition of three auxiliary ASCII files named

- filename.idx containing one line for each index entry,
- filename.ilg containing a log of messages created when, behind the scenes, the idx file is automatically processed through makeindex, and
• filename.ind produced by makeindex and containing the formatted index that is then read by \printindex to create the actual index in the document.

With imakeidx, the index is automatically created and incorporated in the document because an execution of makeindex is inserted automatically at the appropriate point when latex or pdflatex is run. Errors and warnings in that step are compiled in the .ilg file (which should be examined because on-screen display of those glitches may well scroll by to quickly to be comprehended).\[58\]

The above-described process yields an index in the default format. Numerous options can be exploited by adding optional arguments in the \makeindex command. For example, the command

\makeindex [options=-s ind_style, intoc]

stipulates that the style defined by the file ind_style.ist should be invoked and the index should be provided—argument intoc—with an entry in the table of contents. The construction of the style file is fully described in Section 12.4 in The \LaTeX Companion as identified in Section A.18.

To assist in constructing the index, an auxiliary package called showidx, which works both with latex and with pdflatex, can be invoked to print the index entries on each page as marginal notes. This package is invoked simply by including the command

\usepackage{showidx}

in the preamble, though it appears as if this command must be inserted after the \makeindex command described above. Unfortunately, when printing in a 6" × 9" area on 8.5" × 11" paper (letterpaper), the marginal insertions of index entries will extend outside the edges of the page and, in .ps and .pdf files, the portions that are off the page will not be displayed in ghostview or acroread. Conveniently, if the .dvi file is displayed on the screen with xdvi (UNIX) or yap (Windows), those marginal notes will be visible in their entirety, even if they extend beyond the boundaries of the page, though that route is available only if figures are provided in .ps or .eps format. For the purposes of checking the index entries, one workaround is to shrink the area of the page in which printing of the text is allowed. For example, for this document, replacing the commands

\setlength{\textwidth}{6.0truein}
\setlength{\oddsidemargin}{0.5truein}
\setlength{\evensidemargin}{0.0truein}

in the preamble with the commands

\setlength{\textwidth}{4.0truein}
\setlength{\oddsidemargin}{1.0truein}
\setlength{\evensidemargin}{1.0truein}

will shrink the text width and reset the margins so there will be space for the marginal notes. To be sure, the pagination of the text will also change, but at least the marginal notes will be visible, even in the .ps and .pdf files. In different situations, tampering with the above illustrated length parameters and, perhaps also, tampering with the length parameters \marginparwidth (which sets the width of the marginal boxes) and \marginparsep (which sets the space between the text and the boxes) as well as seeking ways to shrink the entire output including marginal notes on each page as a unit may yield fruit.\[59\]

\[58\]With makeidx, a separate explicit processing of the .idx file with makeindex to produce the .ind file and a further pass through latex or pdflatex was necessary to include the index in the document.

\[59\]For example, the -x option to dvips can scale the size of each page output to the .ps file.
A.10 Including Hyperlinks

Properly constructed, a \LaTeX source file can be used to create a PDF document that includes a navigation panel and hyperlinked references for easy viewing with a compatible PDF viewer. Only a few changes need be made to the source files so far described to turn entries in the table of contents, internal references in the body of a document, and page references in an index into hyperlinks to the appropriate points in the document, though these features will appear only in PDF files created via the route described in item 1 in Section A.1.3 above. Specifically,

- the \texttt{\documentclass} statement at the beginning of the file must not specify any particular driver, i.e. should be of the form \texttt{\documentclass{Class}}.
- if used at all, the \texttt{\usepackage{graphicx}} command in the preamble must not include any options specifying a graphics driver
- figures must be provided as PDF files or defined using the package \texttt{tikz}. PostScript descriptions of figures will not be properly rendered.
- the command \texttt{\usepackage{hyperref}}, perhaps with some optional arguments, must be added in the preamble and, to make sure that other included packages do not overwrite redefinitions made by \texttt{hyperref}, should be the last package included.

This edited source file is then processed by \texttt{pdflatex} (Section A.1.3) at least twice to include hyperlinks, and tables of contents, figures, and tables. With these simple additions, a navigation panel in the PDF file containing chapter, section, subsection, and ... subdivisions of the document will be created. Further, all entries in the tables of contents, list of figures, and list of tables, all internal references created with \texttt{\ref} commands, and all index entries will be converted into hot links within the document.

The command \texttt{\usepackage{hyperref}} in the preamble is the minimum necessary to create a hyperlinked PDF file. By default, hot links in the document will be displayed in a red box. To override that default, one can exploit options in the \texttt{\usepackage} command. For example, the statements,

\begin{verbatim}
\usepackage{color}
\definecolor{MyPurple}{rgb}{0.577,0.000,1.000}
\usepackage[colorlinks=true,%
linkcolor={MyPurple},bookmarksnumbered=true,linktocpage=true]{hyperref}
\end{verbatim}

in the preamble of the \LaTeX source file will set the stage for automatic creation of hyperlinks for all references, setting the color of those hyperlinks to \texttt{MyPurple} as specified by the RGB values in the \texttt{\definecolor} statement and removing the enclosing box, stipulating that the bookmarks in the navigation panels should include chapter and section numbers, and stipulating that page numbers rather than section titles should be the linked entries in the Table of Contents.\footnote{The default value of the \texttt{bookmarksnumbered} and \texttt{linktocpage} options is \texttt{false}.} The final PDF file is then produced by processing the source file with \texttt{pdflatex} (Section A.1.3)—at least twice and perhaps three times. To avoid error messages, all auxiliary files hanging over from a previous processing of the source file without hyperlinks through \texttt{latex} or \texttt{pdflatex} should be deleted before processing the file with hyperlinks.\footnote{See item 25 in Section A.17 for ways to achieve this deletion.}

The files created in the user's directory when a properly constructed .\texttt{tex} file with hyperlinks, index, and table of contents is processed through \texttt{pdflatex} to produce a linked PDF file are\footnote{We here assume you are making an index and using \texttt{imakeidx}. The sequence and auxiliary files will be different if you are not making an index and/or if you are using \texttt{makeidx}.}
• (after first pass through \texttt{pdflatex}) \texttt{.aux}, \texttt{.idx}, \texttt{.ilg}, \texttt{.ind}, \texttt{.log}, \texttt{.out}, \texttt{.pdf}, and \texttt{.toc}. The \texttt{.idx}, \texttt{.ilg}, and \texttt{.ind} files will be present only if you are creating an index with \texttt{imakeidx}. The \texttt{.out} file contains information about hyperlinks; the rest are as described in item 1 of Section A.1.3, in Section A.8, and in Section A.9. At this point, the PDF file does not have the navigation panel or the table of contents but it does have the index. All of these files are ASCII text files, though the PDF file may contain some non-printing characters. As such they can all be displayed in an available text editor and understood.

• (after the second pass and, if necessary, the third pass through \texttt{pdflatex}) the same as in the previous bullet, though many of those files have been updated. At this point, the navigation panel, the internal hyperlinks, and the index have all been created.

\textbf{A WARNING}: Links to figures and tables will point to the line containing the caption for the figure or table. Captions for these items are perhaps better placed \emph{above} the item rather than below the item.

\textbf{A CONVENIENCE}: If the commands \texttt{\documentclass} and \texttt{\usepackage\{graphicx\}} are phrased \textit{without} options and the files in all \texttt{\includegraphics} commands are presented \textit{without} file type, then the file processed with \texttt{latex} will look for .ps and .eps graphics files and the file processed with \texttt{pdflatex} will look for .pdf files. All files to be sought must, of course, exist.\footnote{See Section A.11 for a means to convert .ps and .eps files to .pdf and vice versa.} This feature makes it easy to process the same source file both with \texttt{latex} and with \texttt{pdflatex}.

\section*{A.11 Converting .eps and .ps Files to .pdf}

When files describing figures are created, it is often easier to output those files from the creating program as .eps or .ps files rather than as .pdf files. Unfortunately, if you desire to produce the final output with \texttt{pdflatex}, files describing pictures need to be .pdf. PostScript files must therefore be converted into PDF before running \texttt{pdflatex}. Short of asking someone who already knows how to achieve that transfer, rummaging on the web for guidance may be a frustrating experience. Once you find the right tools, however, the process is quite simple. Basically, to convert the files \texttt{figure.eps} and \texttt{figure.ps} to a useful PDF file involves the two steps

1. \texttt{ps2pdf figure.eps} \hspace{1cm} \texttt{ps2pdf figure.ps} \hspace{1cm} (which will yield \texttt{figure.eps.pdf}) \hspace{1cm} (which will yield \texttt{figure.pdf})
2. \texttt{pdfcrop figure.eps.pdf} \hspace{1cm} \texttt{pdfcrop figure.pdf} \hspace{1cm} (which will yield \texttt{figure.eps-crop.pdf}) \hspace{1cm} (which will yield \texttt{figure-crop.pdf})

Here, step 1 creates the .pdf file and step 2 removes extraneous white space around the perimeter of the first-created .pdf file. Second arguments to both \texttt{ps2pdf} and \texttt{pdfcrop}, as in

\begin{verbatim}
ps2pdf figure.eps figure.pdf
pdfcrop figure.pdf figurecrop.pdf
\end{verbatim}

can be used to relieve the awkwardness of the file names. In addition, \texttt{pdfcrop} has a number of options. The statement \texttt{pdfcrop --help} will provide a list of those options on the screen. Sometimes the tight cropping provided by the illustrated commands will clip the edges a bit too closely. The option \texttt{--margins ...}, as in

\begin{verbatim}
pdfcrop --margins 10 figure.pdf figurecrop.pdf
\end{verbatim}
A.11. CONVERTING .EPS AND .PS FILES TO .PDF

will provide a 10-pixel margin of white space around the image. See the help message for further
details.

To simplify the process, the MS DOS batch files listed in Section A.A.1.1, respectively, and
the UNIX Shell scripts listed in Section A.A.2.1, respectively, exploit the second argument for both
ps2pdf and pdfcrop to control the file names and, when executed with statements like

<table>
<thead>
<tr>
<th>MS DOS</th>
<th>UNIX</th>
</tr>
</thead>
<tbody>
<tr>
<td>ceps2pdf figure</td>
<td>./ceps2pdf figure     (for .eps file)</td>
</tr>
<tr>
<td>cps2pdf figure</td>
<td>./cps2pdf figure      (for .ps file)</td>
</tr>
</tbody>
</table>

will (1) leave in the involved directory a cropped PDF file whose name `figure.pdf` differs from that
of the original PostScript in only the file type and (2) remove any temporary files created along the
way.

The above procedure is a bit tedious if you have more than a few files to convert. If you work
in a Command window (Windows) or a Shell window (UNIX), the steps

• Create a temporary directory and create the several files listed in Section A.A.1 (Windows) or
  A.A.2 (UNIX) and the file listed in Section A.A.3 (Windows and UNIX) into that directory.
  These files can be created by direct typing or they can be copied from the directory `$HEAD/tex`.

• Copy all .ps or all .eps files to be converted into that temporary directory. *This step guards
  against disaster should the process to come be run in the initial directory and fail.*

• Create a file containing a list of the names of the .eps or .ps files in the directory. The
  statements

<table>
<thead>
<tr>
<th>MS DOS</th>
<th>UNIX</th>
</tr>
</thead>
<tbody>
<tr>
<td>dir /b *.eps &gt; dir.txt</td>
<td>ls -1 *.eps &gt; dir.txt</td>
</tr>
<tr>
<td>dir /b *.ps &gt; dir.txt</td>
<td>ls -1 *.ps &gt; dir.txt</td>
</tr>
</tbody>
</table>

utilizing the option /b in Windows and the option -1 (one, not el) in UNIX produce the desired
file (filename, including file type).

• Strip the file type by running the python program `ExtractFileName.py` (Section A.A.3.1).
  Here the same code works in both Windows and UNIX. This action will create the file
  `nameonly.txt` containing only the names of the files to be converted.

• Effect the conversions by running the script `rdfile` with statements like

<table>
<thead>
<tr>
<th>MS DOS</th>
<th>UNIX</th>
</tr>
</thead>
<tbody>
<tr>
<td>rdfileeps (for eps files)</td>
<td>./rdfileeps (for eps files)</td>
</tr>
<tr>
<td>rdfileps (for ps files)</td>
<td>./rdfileps (for ps files)</td>
</tr>
</tbody>
</table>

This action will convert each file in turn from .eps or .ps to .pdf and leave no intermediate
files in the directory.

• Delete the *.eps or *.ps files in the temporary directory.

• Move the *.pdf files to the directory from which the .eps and .ps files were moved temporarily,
  leaving both the original .ps and .eps files intact but adding the .pdf files, so either \texttt{latex}
  or \texttt{pdflatex} can then be used to create the final document.

will accomplish that conversion

The reverse process of converting a .pdf file to .ps or .eps format is easier. Specifically, the
statements
A.12 Using Conditional Expressions in \LaTeX

The \LaTeX package \texttt{ifthen} provides a capacity to include text and \LaTeX commands conditionally, i.e., depending on the state of a Boolean flag. If the package is to be used, the preamble of the \LaTeX source file must contain the command

\begin{verbatim}
\usepackage{ifthen}
\end{verbatim}

Once that command has been processed, we must—also in the preamble—then define one or more Boolean flags and set their values with commands like

\begin{verbatim}
newboolean{FlagName}
setboolean{FlagName}{FlagValue}
\end{verbatim}

where \texttt{FlagName} can be any name that does not conflict with names already used and \texttt{FlagValue} will be either \texttt{true} or \texttt{false}. Finally, at the point in the source file where some text is to be included conditionally, we would place the command

\begin{verbatim}
ifthenelse{\boolean{FlagName}}{
 {Text to be inserted if FlagValue is true.}
} {
 {Text to be inserted if FlagValue is false.}
\end{verbatim}

Either text can be null, in which case the opening and closing braces should still appear side by side ({}), and either text can include \LaTeX commands—which will be executed—and can spread over several lines. In particular, the text can include commands to input additional files. Indeed, the customization in \textit{CPSUP} is achieved by including or excluding files depending on the state of several flags. For example, in the preamble of the \LaTeX source file for \textit{CPSUP}, the command

\begin{verbatim}
\usepackage{ifthen}
\end{verbatim}

appears and several flags are defined and set with commands like, for example,

\begin{verbatim}
newboolean{LATEX} \setboolean{LATEX}{true}
\end{verbatim}

Finally, the command

\begin{verbatim}
ifthenelse{\boolean{LATEX}}{
 {\input{FileName} }
} {
 }
\end{verbatim}

is placed at the point where the file containing the \LaTeX portions of this book would be inserted. If the file identified is in the working directory, only its name need be included; otherwise, a full path—absolute or relative—must be specified. At this point, the \LaTeX Appendix will be included if the flag is \texttt{true} and omitted if the flag is \texttt{false}.

More complicated logical operations can be constructed with the operators \texttt{\and}, \texttt{\or}, and \texttt{\not}. For example, if the Boolean flags \texttt{NUMREC} and \texttt{FORTRAN} are defined and set, then the composite statement


\ifthenelse{\boolean{NUMREC} \and \boolean{FORTRAN} }{ \input{FileName} }{ }

would include FileName only if both NUMREC and FORTRAN are true.

In the effort to render files processable with either \texttt{latex} or \texttt{pdflatex} and also to create easily either a printable document or a linked document, the package \texttt{hyperref} must be invoked only when a linked document is desired. In order to avoid manual editing of the source file, one can insert in the preamble the statements\footnote{Inclusion of the package \texttt{ifthen} can be omitted if it had been entered previously for other reasons.}

\begin{verbatim}
\usepackage{ifthen}
\newboolean{PRINT}

\typeout{}\typein[trueorfalse]{true (for print), false (for linked):}
\ifthenelse{\equal{trueorfalse}{true} \or \equal{trueorfalse}{false}}{ \setboolean{PRINT}{trueorfalse} }{\typeout{} \typeout{Must be either "true" or "false". Try again.} \end{document}}
\end{verbatim}

which define the boolean variable PRINT, ask for entry of either \texttt{true} or \texttt{false} at execution time, and terminate execution if an invalid value is entered. Then, the lines

\begin{verbatim}
\usepackage{color}
\ifthenelse{\boolean{PRINT}}{}{\definecolor{MyPurple}{rgb}{0.577,0.000,1.000}
\usepackage[colorlinks=true, linkcolor={MyPurple}]{hyperref}}
\end{verbatim}

also in the preamble, will include the package \texttt{hyperref}, but only if PRINT is false.

One further conditional command included in \texttt{latex} and \texttt{pdflatex} without adding a package facilitates testing whether a file exists. For example, the statement

\begin{verbatim}
\IfFileExists{test.tex}{\input{test.tex}{}\end{document}}
\end{verbatim}

inputs \texttt{test.tex} if it exists and simply moves on to the next statement if it doesn’t exist. This statement is useful if the file to be input is created in the first pass through \texttt{latex} or \texttt{pdflatex} and then read in a subsequent pass. The “false” clause could alternatively be used to print an error message and terminate execution, e.g.,

\begin{verbatim}
\IfFileExists{test.tex}{\input{test.tex}{}\typeout{File not found} \end{document}}
\end{verbatim}

\section*{A.13 Error Messages Generated by \LaTeX\textsuperscript{X}}

However carefully the code is created, sooner or later \LaTeX\textsuperscript{X} will generate an error message, which will have the general form

\begin{verbatim}
! error identification
1. number text read
               text not read
?
\end{verbatim}
The exclamation point is followed by a message that explains the nature of the error. The lower case l (el) is followed (1) by the number of the line in the code at which the problem is detected and (2) the text surrounding the problem. Finally, the question mark prompts for user input. A simple RETURN at the question mark will instruct \TeX to continue to read the file unless a fatal error has occurred (though \TeX may be forced to make assumptions in order to continue); entry of the character x (followed by RETURN) will exit immediately from the program. Many errors are caused by a missing delimiter or by failure to specify a switch to math mode before a mathematical expression or back to text mode after a mathematical expression. All of the reported errors must be identified and fixed (by editing the code) before \TeX will produce the desired finished product.

A full description of the messages that may be produced as well as an enumeration of other user inputs at the ? will be found in Chapter 8 in The \TeX Manual.

\section{The Page Previewer for .dvi Files}

Since \TeX is not wysiwyg, we must use a page previewer to see exactly what our document looks like without printing it out many, many times. The Local Guide describes the page previewer available at your site. UNIX systems usually provide a version of xdvi and Windows systems usually provide a version of yap, but many such previewers exist and, on personal computers especially, the previewer may be accessed by a mouse click in a GUI that also gives access to other components of the \TeX package. To use this previewer, we first create the .dvi file by running our code through \TeX as described in Section A.1. Then we enter a command like

\begin{verbatim}
xdvi filename or yap filename
\end{verbatim}

(The default extension for the input file is .dvi.)

or select an appropriate item from a menu (as described in The Local Guide). Presently, the first page of the document will appear on the screen. Some versions of xdvi display an array of buttons along the right edge of the Xdvi window; other, probably newer, versions replace those buttons with drop-down menus and icons in a toolbar. Clicking ML on any of the various buttons or selecting items from one of the menus effects the associated action. For xdvi and its clones, possible actions include the options shown in Table A.11. In addition, (1) moving the cursor to any point in the window and pressing (and holding) any of the three mouse buttons will generate a magnified version of a region whose size is determined by which button is pushed and (2) some versions of xdvi have a window along the left edge in which you can click ML on a page number to request immediate display of the selected page—though note that these numbers simply count pages from the beginning of the displayed document and may not correspond to the numbers actually printed on the pages in the document.

Beyond the controls made available through buttons and/or menus around the periphery of the Xdvi window, xdvi responds to a number of keystrokes issued when the the cursor is moved into the area displaying text. Some of these are identified in Table A.12; all are described in the xdvi documentation, e.g., the UNIX man page.

In some environments, a GUI may provide an easy way to use the mouse to work on \TeX code, process it repeatedly with \TeX, and examine the changes with an on-screen previewer each time. In the absence of a GUI giving access to all of these features (text editor, \TeX, screen previewer, printer, ...), one convenient strategy for using a text editor and a page previewer simultaneously involves invoking the previewer with a command like

\begin{verbatim}
man xdvi
\end{verbatim}

The point at which \TeX encounters difficulty is conveyed by the downward displacement of the text after that point, though that point is not always where the offense actually occurs.

Yet Another Previewer

Additional information about xdvi on UNIX systems can be found in the on-line help, accessed in many systems by typing the command man xdvi.

If explicit specification of the input file is omitted, some implementations of xdvi will bring up a browser in which the desired .dvi file can be selected. Other versions may open the most recently displayed file, in which case the browser can be invoked by selecting ‘Open’ from the File menu.
Table A.11: Some of the features available through selection from menus (first column) in some versions of \texttt{xdvi}, through clicking ML on buttons in other versions of \texttt{xdvi}. The features actually available by these means may vary from version to version of \texttt{xdvi}.

<table>
<thead>
<tr>
<th>Menu</th>
<th>Button</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>\texttt{File→Reload}</td>
<td>Reread</td>
<td>Rereads current \texttt{.dvi} file.</td>
</tr>
<tr>
<td>\texttt{File→Quit}</td>
<td>Quit</td>
<td>Quits previewer.</td>
</tr>
<tr>
<td>\texttt{Zoom→Shrink by 1}</td>
<td>100%</td>
<td>Displays text at full size (quite large).</td>
</tr>
<tr>
<td>\texttt{Zoom→Shrink by 3}</td>
<td>33%</td>
<td>Displays text at 33% of full size.</td>
</tr>
<tr>
<td>\texttt{Zoom→Shrink by 4}</td>
<td>25%</td>
<td>Displays text at 25% of full size.</td>
</tr>
<tr>
<td>\texttt{Zoom→Shrink by 6}</td>
<td>17%</td>
<td>Displays text at 17% of full size.</td>
</tr>
<tr>
<td>\texttt{Navigate→First Page}</td>
<td>First</td>
<td>Moves to first page.</td>
</tr>
<tr>
<td>\texttt{Navigate→Page-10}</td>
<td>Page-10</td>
<td>Moves to the tenth page before the one displayed.</td>
</tr>
<tr>
<td>\texttt{Navigate→Page-5}</td>
<td>Page-5</td>
<td>Moves to the fifth page before the one displayed.</td>
</tr>
<tr>
<td>\texttt{Navigate→Prev}</td>
<td>Prev</td>
<td>Moves to previous page.</td>
</tr>
<tr>
<td>\texttt{Navigate→Next}</td>
<td>Next</td>
<td>Moves to next page.</td>
</tr>
<tr>
<td>\texttt{Navigate→Page+5}</td>
<td>Page+5</td>
<td>Moves to the fifth page after the one displayed.</td>
</tr>
<tr>
<td>\texttt{Navigate→Page+10}</td>
<td>Page+10</td>
<td>Moves to the tenth page after the one displayed.</td>
</tr>
<tr>
<td>\texttt{Navigate→Last Page}</td>
<td>Last</td>
<td>Moves to last page.</td>
</tr>
<tr>
<td>\texttt{Options→PostScript}</td>
<td>View Page</td>
<td>Allows selection among two or more options, including displaying figures (the initial state) and replacing them with properly sized rectangles.</td>
</tr>
<tr>
<td>\texttt{File→Open}</td>
<td>File</td>
<td>Brings up a browser for selection of new \texttt{.dvi} file.</td>
</tr>
</tbody>
</table>

Table A.12: Keystrokes for controlling \texttt{xdvi}.

<table>
<thead>
<tr>
<th>Keystroke</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>\texttt{s}</td>
<td>Size text to fit window.</td>
</tr>
<tr>
<td>\texttt{ls}</td>
<td>(one-s) Largest text.</td>
</tr>
<tr>
<td>\texttt{ns}</td>
<td>Apply shrink factor \texttt{n}.</td>
</tr>
<tr>
<td>\texttt{g}</td>
<td>Move to last page.</td>
</tr>
<tr>
<td>\texttt{1g}</td>
<td>(one-g) Move to first page.</td>
</tr>
<tr>
<td>\texttt{ng}</td>
<td>Move to (absolute) page \texttt{n}.</td>
</tr>
<tr>
<td>\texttt{n}</td>
<td>Move to next page.</td>
</tr>
<tr>
<td>\texttt{p}</td>
<td>Move to previous page.</td>
</tr>
<tr>
<td>\texttt{nn}</td>
<td>Advance \texttt{n} pages.</td>
</tr>
<tr>
<td>\texttt{np}</td>
<td>Back up \texttt{n} pages.</td>
</tr>
<tr>
<td>\texttt{q}</td>
<td>Exit program.</td>
</tr>
</tbody>
</table>

\texttt{xdvi filename &} (UNIX) \ or \ \texttt{yap filename &} (Windows)

which will start \texttt{xdvi} or \texttt{yap} but detach it from the launching \texttt{Shell} window, leaving the \texttt{Shell} window free for other uses. Then, with the \texttt{.tex} file in a text editor detached from the launching \texttt{Shell} window, we can edit the text, save the edited version, use the \texttt{Shell} window to invoke \LaTeX\ on the edited file, and then simply click ML at an appropriate point in the displayed text to instruct \texttt{xdvi} or \texttt{yap} to reread the \texttt{.dvi} file.\textsuperscript{69} Thus, the effect of each new set of edits can be examined quickly without repeatedly starting and exiting from the previewer.

A.15 The Spell Checker in UNIX\textsuperscript{70}

The programs \texttt{ispell} and \texttt{aspell}, either or both of which may be part of your \LaTeX\ distribution, are common spell checkers. If either is installed at your site (see the \textit{Local Guide}), it can be invoked

\textsuperscript{69}In some environments, rereading an updated file will happen automatically without an explicit user request.

\textsuperscript{70}To the author’s knowledge, no comparable stand-alone spell checker exists for Windows.
APPENDIX A. INTRODUCTION TO \LaTeX

with a command like\footnote{Additional information about \texttt{ispell} or \texttt{aspell} can be found by typing the command \texttt{ispell} with no arguments or the command \texttt{aspell help}. Even more detailed information may be available in the on-line help, accessed in many systems by typing the command \texttt{man ispell} or the command \texttt{man aspell}. Googling \texttt{ispell} or \texttt{aspell} will surely also provide links to detailed descriptions.}

\texttt{ispell filename} or \texttt{aspell check filename}

or perhaps with a mouse click in a GUI. For files with extension .\texttt{tex}, \texttt{ispell} enters its \TeX{} mode and will not identify every \TeX{} or \LaTeX{} command as a misspelled word. Note, however, that these commands are simply ignored by \texttt{ispell} and \texttt{aspell}; their correctness or legitimacy as commands to \LaTeX{} is not assessed.

\section*{A.16 A Sample Document}

The following sample document demonstrates some of the commands explained above and also introduces some useful commands that have not yet been mentioned. (The explanatory comments following the \% sign can be included in the code but have no effect on the output produced.)

\begin{verbatim}
\documentclass{article} % Mandatory command; select
% default 10 point type
\setlength{\textwidth}{6truein} % Set width of text area
\setlength{\oddsidemargin}{0.25truein} % Center text on 8.5'' width
\setlength{\parskip}{6pt plus 1pt minus 1pt} % Specify extra space
% between paragraphs,
% allowing LaTeX to adjust
% space 1 point up or down
\setlength{\parindent}{40pt} % Set paragraph indent

\begin{document} % Mandatory command

\begin{center} % Large, bold, centered title
\{\Large\bf A Sample Document for Your Perusal\}
\end{center}

\begin{flushleft} % Two lines; flush left
{\em Author\}/: J.~Q.~Student \ \% \% \% forces a new line
% Note italic correction
% Tilde prevents extra space
% Note italic correction
% Tilde prevents extra space
% Note italic correction
% Tilde prevents extra space
\end{flushleft}
\begin{flushleft} % Two lines; flush left
{\em Date\}/: \today
\end{flushleft}

In this document we have used the \verb+\setlength+ commands to modify
\LaTeX{}‘s default page setup to make fuller use of an 8.5'\times11'' page.\footnote{\texttt{\footnote{Note the special command \LaTeX{} \verb+\backslash+\texttt{LaTeX}\verb+\LaTeX{}\verb+ provided for the display of the \LaTeX{} ‘‘logo’’.}}}

The title could have been generated using the three commands \verb+\begin{document}\verb+\title{+\verb+\dots+\verb+\author{+\verb+\dots+\verb+\date{+\verb+\dots+\verb+\begin{document}}}+
\end{verbatim}

\texttt{\footnote{Additional information about \texttt{ispell} or \texttt{aspell} can be found by typing the command \texttt{ispell} with no arguments or the command \texttt{aspell help}. Even more detailed information may be available in the on-line help, accessed in many systems by typing the command \texttt{man ispell} or the command \texttt{man aspell}. Googling \texttt{ispell} or \texttt{aspell} will surely also provide links to detailed descriptions.}}
command---i.e., in the preamble---followed by the command \verb++\maketitle+ after the beginning of the document. The command \verb*+\verb*+ ... +? will cause \LaTeX{} to print whatever is between the plus signs (which could be virtually any specified character) exactly as it is. The asterisk is optional, and it makes \LaTeX{} highlight the spaces that occur inside the \verb+verbatim+ environment.

\noindent Notice that a blank line creates a new paragraph. The \verb+\noindent+ command \em{suppresses} the standard paragraph indentation. In this special space we will also demonstrate some math environment operations. Consider the equation

\begin{equation}
\int_{0}^{\infty} e^{-x} \, dx = 1, \label{LATEX:demo}
\end{equation}

where a medium sized space, specified by the command \verb+:+, is put between the integral sign $\int_{0}^{\infty}$, and the integrand and a small space, specified by the command \verb+,:+, is inserted between the integrand and the $dx$ in Eq.~(\ref{LATEX:demo}). The equation

\begin{equation}
\frac{\partial e^{x_0 y^2}}{\partial x_0} = y^2 e^{x_0 y^2}
\end{equation}

demonstrates partial derivatives and fractions. Note the automatic generation and placement of equation numbers in the output. Also note that pages are automatically numbered, though the actual printing of page numbers can be suppressed with the commands \verb+\pagestyle+ and \verb+\thispagestyle+.

The output produced when this short sample file is processed through \LaTeX{} and \texttt{dvips} and then printed is shown in Table A.13, though the page header on that page is not included in the output. Note that, to obtain the proper internal reference to the first equation, this document must be processed twice by \LaTeX{}. The file itself is named \texttt{texsample2.tex} and can be copied from the directory $\HEAD/tex$.

\section*{A.17 Miscellaneous Other Features}

\LaTeX{} has an enormous number of additional features beyond those enumerated above. In particular, be aware of

1. All of the environments listed in Section A.3.4.
2. All of the packages listed in Section A.3.5.
3. The command \texttt{\today}, which returns today's date in the form “month date, year”.
4. The command \texttt{\newcommand}, which permits us to define commands supplementing the standard commands. For example, the commands

\begin{verbatim}
\newcommand{\beq}{\begin{equation}}
\newcommand{\eeq}{\end{equation}}
\end{verbatim}

\begin{verbatim}
\newcommand{\beq}{\begin{equation}}
\newcommand{\eeq}{\end{equation}}
\end{verbatim}
Table A.13: Output produced by the code in Section A.16, except that the equations here are numbered (A.1) and (A.2) rather than (1) and (2) and the footnote is labeled \textsuperscript{a} rather than \textsuperscript{1}. Further, the page number that would appear has been deleted in this display.

### A Sample Document for Your Perusal

**Author:** J. Q. Student  
**Date:** 14 February 2023

In this document we have used the \texttt{\setlength} commands to modify \LaTeX{}’s default page setup to make fuller use of an 8.5” x 11” page.\textsuperscript{a} The title could have been generated using the three commands \texttt{\title{...}}, \texttt{\author{...}}, and \texttt{\date{...}} before the \texttt{\begin{document}} command—i.e., in the preamble—followed by the command \texttt{\maketitle} after the beginning of the document. The command \texttt{\verb*+␣...␣+} will cause \LaTeX{} to print whatever is between the plus signs (which could be virtually any specified character) exactly as it is. The asterisk is optional, and it makes \LaTeX{} highlight the spaces that occur inside the \texttt{verbatim} environment.

Notice that a blank line creates a new paragraph. The \texttt{\noindent} command suppresses the standard paragraph indentation. In this special space we will also demonstrate some math environment operations. Consider the equation

\[ \int_{0}^{\infty} e^{-x} \, dx = 1, \]  
\textsuperscript{(A.1)}

where a medium sized space, specified by the command \texttt{\:)}, is put between the integral sign $\int_{0}^{\infty}$, and the integrand and a small space, specified by the command \texttt{\:}, is inserted between the integrand and the $dx$ in Eq. (A.1). The equation

\[ \frac{\partial e^{x_0 y^2}}{\partial x_0} = y^2 e^{x_0 y^2}, \]  
\textsuperscript{(A.2)}

demonstrates partial derivatives and fractions. Note the automatic generation and placement of equation numbers in the output. Also note that pages are automatically numbered, though the actual printing of page numbers can be suppressed with the commands \texttt{\pagestyle} and \texttt{\thispagestyle}.

\textsuperscript{a}Note the special command \texttt{\LaTeX} provided for the display of the \LaTeX{} “logo”.

---

in the preamble will permit us to type the commands \texttt{\begin{beq}} and \texttt{\end{beq}} rather than the longer forms, measurably simplifying the typing of a document containing many displayed equations. Details will be found in Sections 3.4.1 and C.8.1 in *The \LaTeX{} Manual*.

5. The command \texttt{\renewcommand} for changing commands that already exist. When a command already exists (as, for example, when a command is defined by the selected document class), \texttt{\newcommand} will fail. In those circumstances (as we have seen already in Section A.5), we need the command \texttt{\renewcommand}. As one particular example, the built-in command \texttt{\today} which, by default, formats the date in American style (e.g., April 3, 1938) can be changed to format the date in European style (3 April 1938) with the command

\[
\renewcommand{\today}{\texttt{\number\day\ space \ ifcase\month\ or January% \\
| or February| or March| or April| or May| or June| or July| or August% \\
| or September| or October| or November| or December| fi \ space\number\year}}
\]

The percent signs at the end of the first two lines are not superfluous; they guarantee that \TeX{} will see this command as a single line and hence that extraneous spaces will not appear in
the output for occasional dates. Note also that this modification invokes \TeX's case structure, which is introduced by the \ifcase command and terminated by the \fi command.

6. The command \texttt{\pagestyle{Style}}. Placed in the preamble, this command selects a particular (global) page style \textit{for the entire document}. Recognized styles include \texttt{plain}, \texttt{empty}, \texttt{headings}, and \texttt{myheadings}. Details will be found in Sections 6.1.2 and C.5.3 in \textit{The \TeX\ Manual}.

7. The command \texttt{\thispagestyle{Style}}. Placed at any point, this command selects a particular page style \textit{for the current page}, overriding the global specification for that page alone. Recognized styles are the same as for the command \texttt{\pagestyle}. This command in the form \texttt{\thispagestyle{empty}} is commonly used to suppress page numbering on the first page of a several-page document. Details will be found in Sections 6.1.2 and C.5.3 in \textit{The \TeX\ Manual}.

8. \texttt{ReV\TeX}, which contains files defining the \texttt{aps} document class. These files have been created by the American Institute of Physics (AIP), the American Physical Society (APS), and the Optical Society of America (OAS) and are intended for use in preparing manuscripts for ultimate publication in the journals published by these organizations. Full documentation is contained in \textit{The \REV\TEX\ Input Guide} prepared by the AIP, the APS, and the OAS.

9. The command \texttt{\input}, which simply inputs the file specified in its argument. Details are described in Section 4.4 of \textit{The \TeX\ Manual}.

10. The calligraphic type style for producing upper-case calligraphic letters in math mode. This style can be invoked either with the declaration \texttt{\cal} or the command \texttt{\mathcal}. It is described in Sections 3.3.2, 3.3.8, and C.7.8 of \textit{The \TeX\ Manual}.

11. Methods for placing accents over characters. For example, the command \texttt{"{o}} will produce \text{"{o}}, the command \texttt{\~{n}} will produce \text{\~{n}}, and the command \texttt{\c{c}} will produce \text{\c{c}}. A full listing of the possibilities will be found in Section 3.2.1 and Table 3.1 of \textit{The \TeX\ Manual}.

12. The \texttt{alltt} and \texttt{verbatim} packages, which allow incorporation of computer code by reference to the actual file containing the computer program itself. On the surface, it would appear as if a file containing computer code could be incorporated into a \texttt{\LaTeX} document with the simple command sequence

\begin{verbatim} \input{FileName} \end{verbatim}

The problem comes because the backslash that introduces the command \texttt{\input} will, in the \texttt{verbatim} environment, be treated as an ordinary character and will \textit{not} be recognized as introducing a command. If, instead, we invoke the alternative \texttt{alltt} environment with the \texttt{\LaTeX} commands

\begin{verbatim} \input{FileName} \end{verbatim}

the problem is solved, since \textbackslash, \{, \}, and one or two other characters are treated specially within the \texttt{alltt} environment and the embedded command \texttt{\input} will now be properly recognized as a command to read the specified file into the \texttt{\LaTeX} source stream at this point. This new environment, however, will be available only if the \texttt{alltt} package is explicitly added with the command

\texttt{\usepackage{alltt}}

placed in the preamble to the \texttt{\LaTeX} file.

\footnote{Version 4.1, which was released in final form on 11 August 2010, is compatible with \texttt{\LaTeX2e}. The APS website publish.aps.org/revtex4 \texttt{(no www.)} contains up-to-date information about \texttt{ReV\TeX}4, links to an assortment of manuals (including one titled \textit{Rev\TeX}4.1 Author's Guide}, and a link from which that version can be downloaded. \texttt{ReV\TeX} is automatically included in many standard \texttt{\LaTeX} distributions.
Actually, there is at least one situation in which the `alltt` environment isn’t quite up to the task. If we want to read in a C program that specifies newline characters with `\n`, `\LaTeX` embellished with `alltt` will complain that `\n` is an undefined command. For this case (and, of course, for the others as well), we need instead exploit the `verbatim` package (not environment), which is made available by placing the command

```
\usepackage{verbatim}
```

in the preamble. In particular, this inclusion defines a `command \verbatiminput`, invoked with a statement like

```
\verbatiminput{FileName}
```

When the `verbatim` package is invoked, we also have available a new environment—the `comment` environment, which can be used to “bracket” extended text that one wants to exclude from processing by `\LaTeX`.

13. The ten characters (`\$, \&`, `\{`, `\}`, `%`, `\#`, `\`, `\~`, `\^`, `\`), which have special meanings to `\LaTeX` and will not normally be printed as characters. The first seven of these characters can be printed as characters by preceding the character with a backslash, e.g., to print `\$`, type `$\$` in the code. The last three, however, are trickier because `\~` and `\^` are themselves commands for accents over the following letter and `\` is the command for a new line. These last three characters can, however, be printed by invoking the constructions `\verb+~+`, `\verb+^+`, and `\verb+\+`. Here, the `\verb` command enters verbatim mode, the `immediately following + sign—any character can be used—flags the beginning of the text to be presented in verbatim mode, and the final `+` sign—better, repeat of the first character—flags the end of the text to be presented in verbatim mode. The backslash can also be produced in math mode with the command `$\backslash$`.

In some environments and in footnotes, the command `\verb` is forbidden. Another way to specify the printing of some of these special characters is to use the `\TeX` (not `\LaTeX`) command `\char`. In its raw form, this command simply takes a number, e.g., `\char98`, which specifies the insertion of character 98 from the current font. We can, however, also specify the character by invoking the ‘tic’ operator, which instructs `\TeX` to calculate the numeric code from the character itself. In particular, some of the special characters that cannot be invoked directly can be invoked with the command `\char`. Thus, for example, the command `\char97` will produce the character ‘a’ while the commands `\char`b`, `\char`r`, `\char`\`, and `{\tt \char`\}` will produce characters ‘b’, ‘r’, ‘\’, and ‘\’, respectively. Similarly, the command `{\tt \char`>` will produce ‘>’, which differs in sometimes desirable ways from the character ‘>’ produced by `$\$`.

14. The commands that allow us to create boxes as described in Sections 6.4.3 and C.13.3 in `\LaTeX` in `\LaTeX` Manual. These commands include `\mbox`, `\makebox`, `\fbox`, `\framebox`, `\parbox`, and `\rule` and the `minipage` environment. All of these commands “block” the contents of the box into a structure that `\LaTeX` sees as a single unit. The commands `\fbox` and `\framebox` also place a printed rectangular box around the contents of the box created.

15. The notion of counters. In the off-the-shelf version of `\LaTeX`, each new invocation of the `enumerate` environment starts a counter off again at the beginning. Sometimes, we might wish to have the numbers in a new enumeration pick up from where the numbers in the previous one ended. We should, of course, achieve that objective in a way that does not depend on knowing explicitly what the starting number in the new environment should be. Basically, we have to invent a way to tell `\LaTeX` to remember the value at which it ended and retrieve that

\footnote{The `verbatim` package also redefines the `verbatim` environment in ways, however, of little consequence unless the text in the environment is very extensive. If the `verbatim` package is invoked, one particularly significant change in the `verbatim` environment results in the complete ignoring of any characters following the statement `\end{verbatim}` in the same line. It is best always to place the statement `\end{verbatim}` on a line by itself.}
value when a new environment is entered. The remembered value must be stored in a \TeX variable called a \textit{counter}, which must be defined with the command

\newcounter{hold}

where \texttt{hold}, which cannot contain numeric characters, is the name chosen for the counter. Then, the structure

\begin{enumerate}
\item Text of first item.
\item Text of second item.
\end{enumerate}

... Text not in enumerate environment.

\begin{enumerate}
\setcounter{hold}{\value{enumi}}
\item Text of third item.
\item Text of fourth item.
\end{enumerate}

will achieve the desired end. In the first \texttt{enumerate} environment, the counter \texttt{enumi}, which is used behind the scenes to keep track of the number of items, is initialized to zero. It is then incremented by 1 with each new item. At the end of the environment, \texttt{enumi} stores the number of the last item. The command \texttt{\setcounter} in the first environment saves the current value of \texttt{enumi} in the counter \texttt{hold}, which survives the exit from the environment. Thus, immediately on entry to the second environment, we can use a “reflection” of the first use of \texttt{\setcounter} to restore the value that \texttt{enumi} had reached at the end of the first environment.

16. Parameters influencing placement of “floats” (figures and tables). Sometimes \TeX seems to have a mind of its own with regard to placement of floating objects on the current or later pages. Many of these “problems” can be cleared up by tampering with the default values of the parameters that limit the number of floats that can be put at the top or bottom of the page or, even more, by changing the parameters that stipulate the maximum fraction of a page that can be devoted to floats or the minimum amount of text that must appear on a page. These parameters are enumerated at the end of Section C9.1 in \textit{The \TeX Manual}. For single column presentations, the most important of these parameters are \texttt{\topfraction}, \texttt{\bottomfraction}, \texttt{\textfraction}, and \texttt{\floatpagefraction}. The command \texttt{\renewcommand} must be used to change the values of these parameters.

17. The way to change section, page, figure, table, and footnote “numbering”. By default in the \texttt{article} class, sections, subsections, subsubsections, pages, figures, and footnotes are labeled with Arabic numbers, e.g., Section 3.1.2. The format of that label as determined from the underlying counters \texttt{section}, \texttt{subsection}, \texttt{subsubsection}, \texttt{page}, \texttt{figure}, \texttt{table}, and \texttt{footnote} can be changed. If, for example, we wished the sections to be labeled with upper-case letters, the subsections with lower-case letters, and the subsubsections with arabic numbers (the default), we would simply execute the instructions

\begin{verbatim}
\renewcommand{\thesection}{\Alph{section}.}
\renewcommand{\thesubsection}{\thesection\alph{subsection}.}
\end{verbatim}

in the preamble.\footnote{Note the explicit periods and the inclusion of the section “number” in the definition of the subsection “number”.

Similarly, the commands

\begin{verbatim}
\renewcommand{\thetable}{\Roman{table}}
\renewcommand{\thepage}{\roman{page}}
\end{verbatim}

\footnote{Note the explicit periods and the inclusion of the section “number” in the definition of the subsection “number”.

Similarly, the commands

\begin{verbatim}
\renewcommand{\thetable}{\Roman{table}}
\renewcommand{\thepage}{\roman{page}}
\end{verbatim}
Table A.14: Structure to produce full-width text at top of double-column text.

```latex
\documentclass[twocolumn]{article}
   ... Whatever preamble we want
\begin{document}
\twocolumn
[ \begin{center}
   \\begin{large}\bf Title} \\[12pt] \% Bold title; extra space
   \\begin{large}\bf Author} \\[12pt] \% Bold author; extra space
   \end{center}
\begin{quotation}
   \noindent ... Text of abstract.
   \end{quotation}
]
...
\end{document}
```

will change table “numbering” to upper-case Roman numbers and page “numbering” to lower-case Roman numbers. Essentially, the “numbers” on these components of a document are generated when \LaTeX{} executes the command `\the...`, where `...` stands for the name of the appropriate counter. Redefining this command changes the translation \LaTeX{} applies to the counter when generating the “number”.

18. The way to create a full-width heading and abstract above (and on the same page as) a two-column presentation of text, which exploits an optional argument to the command `\twocolumn` (not the optional argument `twocolumn` to the command `\documentclass`). One format that achieves this end uses the code is shown in Table A.14.\footnote{In more recent versions of \LaTeX{}, the optional argument `[12pt]` to put a bit of extra space after the title and author appears to generate error messages. The “fix” is to replace `[12pt]` with a full command, specifically `\vspace{12pt}`.} If this pattern is to work, there can be no commands that produce output between the command `\begin{document}` and the command `\twocolumn`.

19. The means to change the label in the caption of a figure. By default, \LaTeX{} introduces the caption of the first figure with the label ‘Figure 1:’, incrementing the number automatically with each subsequent figure. The word ‘Figure’ in this caption is the default value of the command `\figurename`, but that portion of the label can be changed with a simple command like

```latex
\renewcommand{\figurename}{Fig.}
```

in the preamble. Changing the colon after the number is harder, since the specification of that punctuation is embedded in the class file in use and is not brought to the outside in a simple command. To change the colon to a period, for example, we need to find the file `article.cls`, search through the file for colons, replace the critical ones with periods, save the file in the default directory with a new name, and then specify it rather than the standard file in the `\documentclass` command at the beginning of the \LaTeX{} source file.\footnote{The file `article.cls` in the version of \LaTeX{} installed at Lawrence University has only two colons that are not in comments; both should be changed.}
20. The means to change the label in the caption of a table. By default, \LaTeX{} introduces the caption of the first table with the label ‘Table 1:’, incrementing the number automatically with each subsequent table. The word ‘Table’ is supplied by the command \texttt{\tablename{}} and can be changed in the way described for \texttt{\figurename{}} at item 19 above. The colon can also be changed as described in the previous item, though changing the colon for figures will change it in the same way for tables.

21. The means to change the label in the title of a chapter. By default, \LaTeX{} introduces the title of the first chapter with the label ‘Chapter 1’ on a line by itself, incrementing the number automatically with each subsequent chapter. The word ‘Chapter’ is supplied by the command \texttt{\chaptername{}} and can be changed in the way described for \texttt{\figurename{}} at item 19 above.

22. The means to change the title above the table of contents and above the index. By default, \LaTeX{} labels the table of contents ‘Contents’ and the index ‘Index’. The word ‘Contents’ is supplied by the command \texttt{\contentsname{}} and the word ‘Index’ is supplied by the command \texttt{\indexname{}}. Each can be changed in the way described for \texttt{\figurename{}} at item 19 above.

23. Using the \texttt{array} environment to create matrices in math mode. As with the \texttt{tabular} environment (see Section A.6), an optional argument specifies the positioning of entries in the columns, ampersands separate entries in each row, and the command \texttt{\\} marks the end of each row. Thus, for example, the \LaTeX{} code

\begin{verbatim}
\[
\left(\begin{array}{cccc}
1 & 2 & 3 & 4 \\
5 & 6 & 7 & 8 \\
9 & 10 & 11 & 12
\end{array}\right)
\]
\end{verbatim}

will produce the display

\[
\begin{pmatrix}
1 & 2 & 3 & 4 \\
5 & 6 & 7 & 8 \\
9 & 10 & 11 & 12
\end{pmatrix}
\]

The entries in each cell can, of course, be much more elaborate than the simple choices made here.

24. The \LaTeX{} package \texttt{needspace}, which facilitates avoiding awkward page breaks conditionally. For example, when a section title occurs close to the bottom of a page, \LaTeX{} may place the title and only a single line of the text before turning the page. Sensible styles would dictate that the section be started on a fresh page. The command \texttt{\newpage} will, of course, achieve that objective. As a document experiences further edits, however, that page break may not any longer be appropriate. Placing the command

\begin{verbatim}
\usepackage{needspace}
\end{verbatim}

in the preamble and then, at the point where a conditional page break might be wise, inserting one of the commands

\begin{verbatim}
\needspace{5\baselineskip} \texttt{or} \Needspace{5\baselineskip}
\end{verbatim}

will result in turning the page, but only if five or fewer lines remain at the bottom of the page. The amount of space left by \texttt{\needspace} will depend some on what penalties are in effect but will usually be close enough to be acceptable; further, \texttt{\needspace} will leave a ragged bottom even if \texttt{\flushbottom} is in effect. The command \texttt{\Needspace} will leave the requested space, will take longer to execute, and will leave a ragged bottom; the command \texttt{\Needspace*} will produce a flush bottom if \texttt{\flushbottom} is in effect. The unit of measure is the size of the length parameter \texttt{\baselineskip}. The 5 in this example can, of course, be whatever number seems appropriate. Note that there is \texttt{no} multiplication sign after the number.
25. The Windows batch command cleanTEX defined by the batch file

\begin{verbatim}
REM delete all TEX intermediate files
\end{verbatim}

and the Unix shell script defined by the file

\begin{verbatim}
#!/bin/bash
# delete all TEX intermediate files
rm -f *.sav *.out *.ind
\end{verbatim}

which facilitate removing all \TeX/\LaTeX\ auxiliary files\(^{77}\) in the directory in which the command or script is executed. These files can be created by direct typing or can be copied from the directory `$HEAD/tex`. The Windows file is executed by the simple statement `cleanTEX` at a command window prompt, though a path to the file may be included if the file is not located in the directory to be purged of \TeX/\LaTeX\ auxiliary files; the UNIX file is executed by the simple statement `./cleanTEX`, though the characters `./` may be replaced by the path to the file if the file is not located in the directory to be purged of \TeX/\LaTeX\ auxiliary files.

26. The Windows batch command `cleanbak` define by the the batch file

\begin{verbatim}
REM Delete .bak files
del /s *.bak
del /s *.sav
\end{verbatim}

and the UNIX shell script `cleanbak` defined by the the file

\begin{verbatim}
#!/bin/bash
# Delete *.bak, *.sav, and *~ files
find . -name '*.bak' -exec rm {} \;
find . -name '*.sav' -exec rm {} \;
find . -name '*~' -exec rm {} \;
\end{verbatim}

which facilitate removing all backup files\(^{78}\) in the directory in which the command or script is executed and in all subdirectories below that directory. These files can be created by direct typing or can be copied from the directory `$HEAD/tex$. The Windows file is executed by the simple statement `cleanbak` at a command window prompt, though a path to the file may be included if the file is not located in the directory at the top of the tree to be purged of backup files; the UNIX file is executed by the simple statement `./cleanbak`, though the characters `./` may be replaced by the path to the file if the file is not located in the directory at the top of the tree to be purged of backup files.

\section*{A.18 References}

As access to the web has expanded, more and more of the information that once was printed is available more readily on the web. Searching for help on specific issues, either on the web as a whole or (perhaps more effectively) more narrowly on the \TeX\ Users Group site \url{www.tug.org} will yield valuable results. Further, when \TeX\ is installed in UNIX and UNIX-based operating systems (which includes Mac computers), the Shell command `man`, e.g. `man hyperref`, may bring up a

\footnote{Files with file types `.aux`, `.log`, `.dvi`, `.toc`, `.ilg`, `.idx`, `.bak`, `.sav`, `.out`, and `.ind.`}

\footnote{Files with file types `.bak` and `.sav` and files for which the last character in the name is `~`.}
Numerous books have also been written for a variety of audiences and with objectives ranging from general discussions to very specific focus on particular tasks. Among the more common books are the following:


\textit{Math Into \LaTeX} (Third Edition), George Grätzer (Birkäuser, Boston) [ISBN 0-8176-4131-9 or 978-0201433111, 2000]

\textit{The \LaTeX \ Web Companion Integrating \TeX, HTML, and XML}, Michael Goossens, Sebastian Rahtz, Eitan M. Gurari, and Ross Moore (Pearson Education, Inc.) [ISBN 0201433117 or 978-0201433111, 1999]

Additional books will likely surface in a search for \LaTeX \ on the Amazon or Barnes and Noble websites, though that search may also generate a number of hits for documents about rubber.

\section*{A.19 Exercises}

\subsection*{A.1.} Study carefully the \LaTeX \ code presented in Section A.16 and the resulting output in Table A.13, making sure you understand both the syntax of each command and the effect it produces in the output.

\subsection*{A.2.} Use \LaTeX \ to write a letter to a friend. Format your letter so that it has

- a centered block at the top containing your name and address (\texttt{center} environment);
- a right-justified date (using \texttt{\today} and either \texttt{\hfill} or the \texttt{flushright} environment);
- a left-justified block containing an inside address (\texttt{flushleft} environment);
- a left-justified salutation;
- the body of the letter, containing more than one paragraph;
- a closing (e.g., Sincerely, With love, …) positioned 3.5” from the left margin; and
APPENDIX A. INTRODUCTION TO \TeX

\begin{itemize}
  \item your name, spaced far enough below the closing to allow for your signature and aligned with the closing.
\end{itemize}

Suppress page numbers altogether on this letter. Include both your \TeX code and the processed output in what you submit as a solution to this exercise.

A.3. Examine one (or more) of the templates listed in Table A.6 and, in a document produced with \TeX, explain the function of each command it contains. Include both your \TeX code and the processed output in what you submit as a solution to this exercise.

A.4. The folder $\textsf{\$HEAD/tex}$ contains the three files radio.ps, radio.eps, and radio.pdf, each of which contains a description of a graph showing the number of nuclei of each of three species $A$, $B$, and $C$ as a function of time as the 1000 nuclei of $A$ initially present decay radioactively first to $B$ and then to $C$. The equations describing the system are

\begin{align*}
\frac{dA}{dt} &= -k_A A; \quad \frac{dB}{dt} = k_A A - k_B B; \quad \frac{dC}{dt} = k_B B
\end{align*}

and the graph shows the solution of these equations when the initial conditions are $A(0) = 1000$, $B(0) = C(0) = 0$ for the case $k_A = k_B = 0.1$. Prepare a document containing this figure, the differential equations, the initial conditions, and a brief description of the graph in English. Be sure that your text includes explicit references to the figure and the equations. You might also contrive to include a footnote somewhere. Include both your \TeX code and the processed output in what you submit as a solution to this exercise. A \textit{Crutch:} The file $\textsf{\$HEAD/tex/sampledoc.tex}$ provides a start on the creation of a suitable source file for this exercise.

A.5. Generate \TeX code to produce the memo

\begin{center}
\textbf{M E M O R A N D U M}
\end{center}

\textit{Current date}

\hspace{1cm} \textbf{TO:} \hspace{2cm} Insert name of recipient here.

\hspace{1cm} \textbf{FROM:} \hspace{2cm} Insert name of sender here.

\hspace{1cm} \textbf{SUBJECT:} \hspace{2cm} Insert topic here.

\textbf{MESSAGE:}

Insert message here.

Make sure that the date placed in the memo is the date the memo was processed (i.e., use the command $\texttt{\today}$). Include both your \TeX code and the processed output in what you submit as a solution to this exercise. \textit{Suggestion:} You may want to save this template in a file, say memo\_template.tex, so you have it available as a starting point for the multitude of memos you will subsequently write.

A.6. In connection with a grant you received from the XYZ Foundation, you need periodically to submit a progress report. The format and content of that report are dictated by the Foundation, and the report is created by filling in the information indicated in italics in the template on page 503. Create a \TeX source file that you can save and use repeatedly to facilitate generating each required report, i.e., create a \TeX source file that, when processed, will produce the output shown below. Pay meticulous attention to the spacing in the heading and to the spacing and the rulings in the table.
PROGRESS REPORT to XYZ FOUNDATION

Due date of report

Grant Number: Grant number
Date of Award: Date of award

Title of Award: Title of award

Principal Investigator (PI): Name of PI

Investigator’s Institution: Name and address of institution

1. Please describe progress made since last report:
   Insert response.

2. Please list talks given since last report, including title and venue:
   Insert response.

3. Please give full citations for each publication since last report:
   Insert response.

4. Please describe any unanticipated difficulties encountered:
   Insert response.

5. Please identify any individuals beyond the PI who have contributed more than ten hours per week to the project during the past time period:
   Insert response.

6. Please summarize expenses since the last report:

<table>
<thead>
<tr>
<th>Category</th>
<th>Amount</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original grant</td>
<td></td>
<td>$xxx,xxx</td>
</tr>
<tr>
<td>Total expenses reported last time</td>
<td></td>
<td>$xxx,xxx</td>
</tr>
<tr>
<td>Available funds at start of current period</td>
<td></td>
<td>$xxx,xxx</td>
</tr>
</tbody>
</table>

   Expenses in current period:
   
   Salaries                               | $xxx,xxx |
   Fringe benefits                        | $xxx,xxx |
   Equipment                              | $xxx,xxx |
   Supplies                               | $xxx,xxx |
   Travel                                 | $xxx,xxx |
   Page charges                           | $xxx,xxx |
   Other: Identify                         | $xxx,xxx |

   Total for current period               | $xxx,xxx |

   AVAILABLE FUNDS FOR REMAINDER OF PROJECT | $xxx,xxx |

Signed: ______________________________  Date: ______________________________

Typed Name of PI

Template to be reproduced in Exercise A.6.
A.A Listings

This section provides listings of the several batch files/shell scripts and the one PYTHON program involved in converting PostScript files to PDF files. Their use is described in Section A.8, and the files themselves reside in the directory $HEAD/tex.

A.A.1 ... for Windows

A.A.1.1 Batch File cept2pdf.bat

REM cept2pdf.bat (Convert EPS to PDF in Windows)
echo off
ps2pdf %1.eps tmp.pdf
pdfcrop tmp.pdf %1.pdf
del tmp.pdf

A.A.1.2 Batch File cps2pdf.bat

REM cps2pdf.bat (Convert PS to PDF in Windows)
echo off
ps2pdf %1.ps tmp.pdf
pdfcrop tmp.pdf %1.pdf
del tmp.pdf

A.A.1.3 Batch File rdfileeps.bat

REM rdfile.bat (Convert each file in nameonly.txt to PDF in Windows)
del dir.txt
for /f %%a in (nameonly.txt) do (
    ps2pdf %%a.eps tmp.pdf
    pdfcrop tmp.pdf %%a.pdf
)
del tmp.pdf
del nameonly.txt

A.A.1.4 Batch File rdfileps.bat

REM rdfileps.bat (Convert each file in nameonly.txt to PDF in Windows)
del dir.txt
for /f %%a in (nameonly.txt) do (
    ps2pdf %%a.ps tmp.pdf
    pdfcrop tmp.pdf %%a.pdf
)
del tmp.pdf
del nameonly.txt
A.A.2 ... for UNIX

A.A.2.1 Shell Script ceps2pdf

#!/bin/bash
# ceps2pdf (Convert EPS to PDF in UNIX)
filename=$1
ps2pdf $filename.eps tmp.pdf
dfcrop tmp.pdf $filename.pdf
rm -f tmp.pdf

A.A.2.2 Shell Script cps2pdf

#!/bin/bash
# cps2pdf (Convert PS to PDF in UNIX)
ps2pdf $1.ps tmp.pdf
dfcrop tmp.pdf $1.pdf
rm -f tmp.pdf

A.A.2.3 Shell Script rdfileeps

#!/bin/bash
# rdfileeps (Convert each file in nameonly.txt to PDF in UNIX)
cat nameonly.txt | while read filename
do
    ps2pdf $filename.eps tmp.pdf
dfcrop tmp.pdf $filename.pdf
    rm -f tmp.pdf
done
rm -f dir.txt
rm -f nameonly.txt

A.A.2.4 Shell Script rdfileps

#!/bin/bash
# rdfileps (Convert each file in nameonly.txt to PDF in UNIX)
cat nameonly.txt | while read filename
do
    ps2pdf $filename.ps tmp.pdf
dfcrop tmp.pdf $filename.pdf
    rm -f tmp.pdf
done
rm -f dir.txt
rm -f nameonly.txt
A.A.3  ... for Windows and UNIX

A.A.3.1 Python Script ExtractFileName.py

# David Cook
# ExtractFileName.py
# Strip file names in dir.txt and store results in nameonly.txt

# Open and read in existing file from execution of dir /b or ls -1
stream = open("dir.txt","r")
all_lines = stream.readlines()
stream.close()

current_line = "\n"
current_heading = ""

# Open a file for writing out
outfile = open("nameonly.txt","w")

# Loop over each line of the 'all_lines' variable, stripping off the file type, leaving only the name to be written to the output file.
for line in all_lines:
    lw = (line.split(". ")[0]).strip()
    outfile.write(lw +"\n")
outfile.close()
Appendix Z

Contacting Software Vendors

Note: Regardless of which components are included and which omitted in this version of *Computation and Problem Solving in Undergraduate Physics*, the information in this Appendix is that from the assemblage containing all components.

In this appendix, we present information to help interested individuals contact the vendors of software referred to at various points in this book. The information in this appendix was accurate as of 16 April 2018, but no guarantee can be made that it will be accurate forever into the future.

**IDL®**

Harris Geospatial Solutions
385 Interlochen Crescent
Broomfield, CO 80021 USA
Voice: 303-786-9900
FAX: 303-786-9909
E-mail: geospatal@harris.com
Web: www.harrisgeospatial.com

In addition, links to numerous third-party contributions of IDL routines can be found by Googling ‘IDL routines’. Among the most prominent of the sites that will emerge points you to the IDL Astronomy User’s Library maintained at the Goddard Space Flight Center and accessible from the URL idlastro.gsfc.nasa.gov.

**LAPACK**

LAPACK is a large FORTRAN 90 package which implements numerous algorithms to accomplish various tasks in linear algebra. Full information is available at [www.netlib.org/lapack](http://www.netlib.org/lapack). The source code is in the public domain. Instructions for downloading the package are included at the referenced website. Those downloads include the routines in BLAS (Basic Linear Algebra Subprograms), which are invoked by routines in LAPACK. The development of the package was supported by the National Science Foundation and the Department of Energy, and maintenance has been supported for many years by MathWorks and Intel. (Netlib is a large repository of programs maintained at Oak Ridge National Laboratory.)

**\LaTeX – see \TeX**

**LSODE**

The ODE solver LSODE is one component in a large package of ODE solvers originating in the Computing and Mathematics Research Division of the Lawrence Livermore National Laboratory. The full package is called ODEPACK, public-domain software written in FORTRAN by Alan C. Hindmarsh and others. Information about compiled solvers, numerous example programs, and source code for the package are available for download from links at [computing.llnl.gov/projects/odepack](http://computing.llnl.gov/projects/odepack) and [computing.llnl.gov/projects/odepack/software](http://computing.llnl.gov/projects/odepack/software). Additional information and downloads may be found from links at [www.netlib.org/odepack](http://www.netlib.org/odepack). In particular, the two text files opkd-sum and
Appendix Z. Contacting Software Vendors

opks-sum at this URL provide a description of the single- and double-precision components in the package. (Netlib is a large repository of programs maintained at Oak Ridge National Laboratory.)

Mac\TeX

Mac\TeX\ is a shareware implementation of the \TeX/\LaTeX\ system for Macintosh computers. Information is available at the URL \url{www.tug.org/mactex}.

MAXIMA

Originally called MACSYMA, this first of the computer algebra systems was developed at MIT and supported from its origin in 1968 until 1982 by MIT, NASA, ONR, and DOE. In 1982, MACSYMA became a commercial product that was further developed and remained available until about 1999. The 1982 MIT version remained available as DOE MACSYMA but was released in 1999 to a group that continues to develop and maintain the program, changing its name to MAXIMA, whose website is at the URL \url{http://maxima.sourceforge.net/}. MAXIMA is freely available for a wide variety of platforms.

MAPLE®

\begin{center}
\begin{tabular}{ll}
Waterloo Maple, Inc. & Voice: 800-267-6583 \\
615 Kumpf Drive & FAX: 519-747-5284 \\
Waterloo, Ontario & E-mail: info@maplesoft.com \\
CANADA, N2V 1K8 & Web: www.maplesoft.com \\
\end{tabular}
\end{center}

MATHEMATICA®

\begin{center}
\begin{tabular}{ll}
100 Trade Center Drive & FAX: 217-398-0747 \\
Champaign, IL 61820-7237 USA & E-mail: info@wolfram.com \\
& Web: www.wolfram.com \\
\end{tabular}
\end{center}

MATLAB®

\begin{center}
\begin{tabular}{ll}
The MathWorks, Inc. & Voice: 508-647-7000 \\
3 Apple Hill Drive & FAX: 508-647-7101 \\
Natick, MA 01760-2098 USA & E-mail: info@mathworks.com \\
& Web: www.mathworks.com \\
\end{tabular}
\end{center}

MiK\TeX

MiK\TeX\ is an implementation of the \TeX/\LaTeX\ system for Windows computers. Information is available at the URL \url{miktex.org}. MiK\TeX\ is also available from CTAN, the \TeX/\LaTeX\ distribution network accessible from links at \url{www.tug.org}.

MUDPACK

The freely-available package of FORTRAN PDE solvers called MUDPACK, written by John C. Adams and others, originated at the National Center for Atmospheric Research (NCAR). The main web page for information about this FORTRAN package, and many others, is \url{www2.cisl.ucar.edu/research-software/software}. MUDPACK can be downloaded from the webpage \url{github.com/NCAR/NCAR-Classic-Libraries-for-Geophysics}. While this package is no longer under active development, it remains available and useful.

Numerical Algorithms Library

The Numerical Algorithms Library (NAG library) is a large commercially available library of C and Fortran subroutines/subprograms implementing a wide assortment of numerical and statistical algorithms.

\begin{center}
\begin{tabular}{ll}
Numerical Algorithms Group, Inc. & Voice: 630-971-2337 \\
801 Warrenville Road & FAX: 630-971-2706 \\
Suite 185 & E-mail: infodesk@nag.com \\
Lisle, IL 60532-4332 USA & Web: www.nag.com \\
\end{tabular}
\end{center}
NUMERICAL RECIPES

The vendor of Numerical Recipes is reluctant to provide detailed contact information, preferring that potential customers deal with them through forms on their website at www.numerical.recipes (without a ‘.com’). The current version and many of the past versions, some of which include languages no longer being updated, are available for download from this site.

OCTAVE

Available under the terms of the GNU General Public License as published by the Free Software Foundation, OCTAVE is an array processing program whose syntax is similar to that of MATLAB. Information about the program can be found at the site octave.org.

ODEPACK – see LSODE

OzTEX

OzTEX is a shareware implementation of the \TeX/La\TeX system for Macintosh computers. Information is available at the URL www.trevorrow.com/oztex. OzTEX is also available from CTAN, the \TeX/La\TeX distribution network accessible from links at www.tug.org.

PYTHON/NUMPY/MATPLOTLIB/PLOTLY/MAYAVI/SCIKIT-IMAGE

PYTHON, a high-level programming language created by Guido van Rossum, was first released in 1991. In execution, PYTHON programs are interpreted, not compiled. Information and free downloads are available at the URL www.python.org. Information about various add-on modules can be found as follows:

- The module numpy, which adds numerous mathematical capabilities (arrays, matrices, mathematical functions, ...) to PYTHON, is described at the URL www.numpy.org.
- The module matplotlib, which adds numerous 2D plotting capabilities and a few 3D plotting capabilities to PYTHON, is described at the URL matplotlib.org.
- The module plotly, which provides several 3D plotting capabilities, is described at the URL plot.ly.
- The module mayavi, which provides numerous capabilities for 3D data visualisation and plotting, is described at the URL docs.enthought.com/mayavi/mayavi.
- The module scikit-image, which includes scimage and provides image processing capabilities, is described at the URL scikit-image.org.

There are numerous distributions of PYTHON. A popular distribution that automatically includes numerous modules that might not be included in many other distributions is described at www.anaconda.com.

\TeX/La\TeX/dvips/graphicx/makeindex/xdvi/dvips/...

The primary site for information (history, current plans, downloads, ...) for \TeX, \LaTeX, and numerous other publicly available components of \TeX and its derivatives is the web site of the \TeX Users’ Group (TUG), www.tug.org. This organization maintains CTAN (the Comprehensive \TeX Archive Network), which has a handful of backbone machines around the world and a number of mirror sites, from any of which an enormous number of files associated with the \TeX/La\TeX system can be downloaded.

\TeXlive

\TeXlive is among the newer implementations of the \TeX/La\TeX system for all platforms. Information about this distribution and instructions for downloading and installing it are available from the URL www.tug.org/texlive.
APPENDIX Z. CONTACTING SOFTWARE VENDORS

TGIF

TGIF (pronounced T-G-I-F) is a versatile program for creating two-dimensional drawings. The program is Xlib-based and interactive, and it runs under X11 on LINUX and UNIX platforms (including MAC OS X and cygwin on Windows). Information (brief history, licensing understandings, instructions for downloading, ...) about TGIF is available from the URL bourbon.usc.edu/tgif/.

Winedt

Winedt is an inexpensive ($40.00 per student user, $60 per educational user) editor for Intel-based machines running one or another version of Microsoft Windows. It has particular features that make it especially suitable for creating source files for TeX/\LaTeX. Information about licensing and downloads can be found at the URL www.winedt.com. Winedt is also available from CTAN, the \TeX/\LaTeX distribution network accessible from links at www.tug.org.

Xemacs

This flexible, open source text editor is available from a variety of sources and is protected under the terms of the GNU Public License. Downloads and information about the program can be found at the URL xemacs.sourceforge.net.

Xv

Written, maintained, and copyrighted by John Bradley, xv is an interactive image manipulation program that runs in X-windows. Downloads, information about the program, and information about licensing and registering is available at the URL www.trilon.com.
Index

The user of this index should be aware that not only textual discussions but also some of the problems are indexed. Page numbers displayed in Roman type refer to textual discussions; page numbers displayed in Italic type refer to problems.

Symbols

$\text{HEAD}$, 11, 15, 125, 221, 244, 266, 284, 323, 324, 374, 459
$\text{IDLHEAD}$, 11
$\text{MATHEMHEAD}$, 137
$\text{NRHEAD}$, 11

A

aborting execution
in MATHEMATICA, 92
in OCTAVE, 16
accuracy
in evaluating integrals, 279–281, 284–287
in finding roots, 322–323
in solutions via FEM, 435–437
in solving ODEs, 218–219, 228, 232
with conservation laws, 219, 232
in solving PDEs, 402–404, 408–410, 416
acroread, 464
action statements, 151, 154
Adams/Bashforth/Moulton (ABM) methods,
220–221
adaptive methods
for integrals, 281
for ODEs, 219–220
adding text to graphics, see placing text in
graphics
algorithms
definition, 151
for $a^2 + b^2 = c^2$, 177
for adding, 177
for bowling, 175
for controlling loops, 160
for counting, 160, 161
for exchange, 160
for finding extremes, 161, 162
for integration, 276–284
for Laplace’s equation, 162
for ODEs, 212–221
for parking lot, 175
for root finding, 320–323
for searching, 162, 175
for sorting, 177
for summing, 161
animation
in MATHEMATICA, 133
in OCTAVE, 69–70, 400–402, 406–408
arithmetic operators
in MATHEMATICA, 95
in OCTAVE, 22
array indices
in MATHEMATICA, 100
in OCTAVE, 19, 168
in pseudocode, 153
arrays
for data storage, 164
in MATHEMATICA, 99
in OCTAVE, 19–21
in pseudocode, 153
scalar, see scalar fields
vector, see vector fields
ASCII character codes, 5
ASCII text files, 8
aspell, 491–492
assigning values to variables
in MATHEMATICA, 96–97
in OCTAVE, 18
in pseudocode, 154

B

backsubstitution, 340
backward difference, 370
Bessel functions, 305, 348
binary files, 8
binary number system, 4–5
Biot-Savart law, 269
bisection, see root finding
bit, 4–5
black body radiation, 144, 304
blocking statements, see also control structures
  in MATHEMATICA, 123
  in OCTAVE, 33, 34
  in pseudocode, 156, 166

Boolean values, see logical expressions

Boolean variables, see logical expressions

boundary conditions
  Dirichlet, 355, 369, 371–372
  mixed, 355, 369, 371–372
  Neumann, 355, 369
  of first kind, see boundary conditions, Dirichlet
  of second kind, see boundary conditions, Neumann
  of third kind, see boundary conditions, mixed

boundary value problems, 185, 196, 206–209, 351, 352, 355

byte, 5–6

C

C, 1

cardioid, 45, 302

carrying capacity, 190, 201–202

Cartesian coordinates
  in MATHEMATICA, 149

case sensitivity
  in MATHEMATICA, 92
  in OCTAVE, 16
  in pseudocode, 152

cells
  in a MATHEMATICA notebook, 101

center of mass
  definition of, 256
  of cardioid, 302
  of semicircular plate, 257, 265

central difference, 370

central processing unit, 3

chaos
  Lorenz system, 252

character codes, 4–5

character strings
  in MATHEMATICA, 94
  in OCTAVE, 17

characteristic equation, 310

charging capacitor, 83

classification
  of boundary conditions, 355–356
  of integrals, 255
  of ODEs, 185
  of PDEs, 355

closing notebook
  in MATHEMATICA, 93

color
  in OCTAVE, see also OCTAVE properties,
    color, facecolor, edgecolor, linecolor, 27, 29

combining statements on one line

in MATHEMATICA, 92
in OCTAVE, 16, 33, 34
in pseudocode, 156

command files
  in MATHEMATICA, see MATHEMATICA command files

  in OCTAVE, see OCTAVE scripts

command line editing
  in MATHEMATICA, 101
  in OCTAVE, 16

command/function duality
  in OCTAVE, 19, 21

comments, see also documentation
  in \LaTeX, 492
  in MATHEMATICA, 125
  in OCTAVE, 16, 38, 41
  in pseudocode, 157, 160

common storage
  in OCTAVE, see global variables

compiler directives, 153
  in pseudocode, see pseudocode directives

compiling source file, see Local Guide, 152

complex numbers
  in MATHEMATICA, 105–107
  in OCTAVE, 19

Compton scattering, 84

conditions, see logical conditions

conservation of mass, 364–365

constants
  in MATHEMATICA, see MATHEMATICA constants

  in OCTAVE, see OCTAVE system variables

contact information, 507–510

continuing statement on second line
  in MATHEMATICA, 92
  in OCTAVE, 16
  in pseudocode, 156

control characters, 4

control structures
  in MATHEMATICA, see MATHEMATICA control structures
  in OCTAVE, see OCTAVE control structures

  in pseudocode, see pseudocode control structures

convergence criteria, 164, 218–219, 279–281, 322–323

convergence error, 372–373, 399

convert .eps/.ps to .pdf, 486–487

convert .pdf to .eps/.ps, 487

Cornu spiral, 260, 292–294

coupled oscillators, 192–193, 205–206, 247

CPU, see central processing unit

Cramer’s rule, 340
**INDEX**

<table>
<thead>
<tr>
<th>Term</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>cross product</td>
<td>125–126, 150</td>
</tr>
<tr>
<td>in MATHEMATICA</td>
<td>125–126, 150</td>
</tr>
<tr>
<td>in OCTAVE</td>
<td>24, 82</td>
</tr>
<tr>
<td>current directory, see also default directory</td>
<td>47, 50, 77–78</td>
</tr>
<tr>
<td>curve fitting, see least squares fitting to parabola</td>
<td>148, 350</td>
</tr>
<tr>
<td>customizing</td>
<td>135</td>
</tr>
<tr>
<td>in MATHEMATICA</td>
<td>135</td>
</tr>
<tr>
<td>in OCTAVE</td>
<td>78–79</td>
</tr>
<tr>
<td>cylindrical coordinates</td>
<td>150</td>
</tr>
<tr>
<td>in MATHEMATICA</td>
<td>150</td>
</tr>
</tbody>
</table>

**D**

- data analysis, 259
- data files
  - freefall.dat, 350
  - irrad.octave.dat, 169, 174
  - ldrum.dat, 69
  - pdens.octave.dat, 170
  - quadpole.octave.dat, 174
  - radio.dat, 35, 37, 38
  - radata.dat, 47
  - ruchardt.dat, 352
  - wavedg.octave.dat, 173
- data types
  - floating point, 6–7
  - integer, 6–7
  - string, 6–7
- data typing
  - explicit, 153
  - implicit, 153
  - in OCTAVE, 17
  - in pseudocode, 152–153, 155–156
- default directory, see also current directory, 9
  - in MATHEMATICA, 124, 131, 132, 136–137
  - in OCTAVE, 30–32, 227, 234, 289
- defining ODEs
  - for MATHEMATICA, 115, 199, 201, 202, 205, 206, 245
  - for OCTAVE, 224–225
- deleting functions
  - in MATHEMATICA, 98
- deleting variables
  - in MATHEMATICA, 97, 98, 109
  - in OCTAVE, 21, 33, 34
- demos
  - in MATHEMATICA, 91, 93
- diffraction patterns, see also interference patterns, 168, 306, 348, 349
  - of square aperture, 51–58
- diffusion equation
  - in 1D, 360–361
  - in OCTAVE, 404–410
  - in 2D, 393–396
  - in 3D, 361
- dimensionless casting
  - for plotting, 85

- of integrals, 257–259, 261, 269, 301, 304
- of ODEs, 118–119, 186–191, 193–195, 201–202, 309, 311
- direct algorithm, 372
- discretization
  - of Laplace’s equation, 163
  - of ODEs, 370–372
  - of PDEs, 356, 377
- discretization error, 372–373, 377, 399
- displaying data on screen
  - in MATHEMATICA, 92
  - in OCTAVE, 19–20
  - in pseudocode, 154
- dividing statement between lines, see continuing statement on second line
documentation, see also comments
  - in MATHEMATICA notebooks, 134–135
  - in OCTAVE m-files, 41
dot product
  - in MATHEMATICA, 125–126, 150
  - in OCTAVE, 23, 81
double factorial, 12
drawing space curves
  - with MATHEMATICA, 138–139
  - with OCTAVE, 73–74
dvipdfm, 464
dvips, 463, 481
dynamic allocation of memory, 153

**E**

- elastic swing, 253
electric field, 260
  - in wave guide, 66–67, 170–173
  - of dipole, 180
  - of quadrupole, 173–174
  - of relativistic particle, 85
electrostatic potential, 260
  - of charged disks, 83, 84, 301
  - of charged rings, 260–261, 295, 299, 305
  - of finite line charge, 266–268
  - of spherically symmetric distribution, 301
  - elemental equation in finite element analysis, 382
  - elliptic integrals, 148, 259, 305

- em dash, 472
- em space, 469
- en dash, 472
- en space, 469
- energy levels
  - in quantum oscillator, see harmonic oscillator
  - in quantum well, 209, 313, 330–332, 336–339, 348
- equation of continuity, 364–365
equilibrium points, 308, 316–317
equivalent resistance, 350
- error function, 260, 262, 284, 289, 291, 292, 298–299, 304
- error messages
INDEX

Euler’s method, 212–214, 221–224
improved, see improved Euler method
evaluating integrals
  with MATHEMATICA, 112, 263–275, 296–299
  with OCTAVE, 284–296
exchange algorithm, 160
executing program, see Local Guide
exiting
  from MATHEMATICA, 93
  from OCTAVE, 16
expressing infinity
  in MATHEMATICA, 264
expressions
  in MATHEMATICA, 94–96

F
Fabry-Perot interferometer, 83
factorial, 12, 102
Fibonacci numbers, 82
field mapping, 87, 88
finding eigenvalues/eigenvectors
  with MATHEMATICA, 110–111, 145, 146
  with OCTAVE, 41–42, 82
finding roots, see root finding
finite difference approximations
  for derivatives, 177
  for PDEs, 356
finite difference methods
  error in, 372–373
  in 1D, 370–377
  in 2D, 391–416
  with OCTAVE, 373–377, 399–416
finite element approximations
  for PDEs, 356
finite element methods
  assembling equations, 377
    in 1D, 377–391
    in 2D, 416–437
  incorporating boundary conditions, 377
    with OCTAVE, 386–391, 427–437
fitting data, see curve fitting, least squares fitting
floating point number
  double precision, 7
  single precision, 6
fluid dynamics, 303
fonts, see also Greek letters
  in OCTAVE, 71–73
formatting data files, 164–165
formatting output
  in OCTAVE, 31
  in pseudocode, 154
FORTRAN, 1
forward difference, 370
four-slit interference pattern, 85
Fourier series, 263, 273–275, 302
Fraunhofer diffraction, 168–169
Fresnel integrals, see also Cornu spiral, 260, 306
  functions
    in MATHEMATICA, see MATHEMATICA functions
    in OCTAVE, see OCTAVE functions
typical available, 154

G
Galerkin method, 377, 379–382
gamma function, 12–13
Gauss-Jordan elimination, 377
Gauss-Kronrod rule, see quadrature
Gaussian distribution, 259, 304
Gaussian elimination, 340–341, 377
Gaussian integration, see quadrature
gedit, 1
ghostview, 464
global positioning, 349
global truncation error, see truncation error, global
global variables
  in OCTAVE, 39–41, 295
golden ratio, 347
gravitational force, 194
Greek letters, see also fonts
  in \LaTeX, 472–473
  in MATHEMATICA, 94
  in OCTAVE, 53, 72

H
handles
  in OCTAVE, 29–30, 64–65, 72, 74–75
hard copy
  from MATHEMATICA, 129–132
  from OCTAVE, 49–50
Helmholtz coil, 268–272
Helmholtz equation, 362
help, see on-line help
hexadecimal number system, 5
high-pass filter, 47–49
high-resolution graphing
  in MATHEMATICA, 126–129
  in OCTAVE, 25–28, 43–49, 51–69
Hooke’s law, 309
hyperlinks, 485–486

I
IDL, 1
IDL contact information, 507
improved Euler method, 214–215, 223
indices in \LaTeX, 483
indicial equation, 209
inert commands
INDEX

in MATHEMATICA, see suppressing evaluation in MATHEMATICA
inhomogeneities in linear equations, 340
initial value problems, 185, 355
initialization files
  in MATHEMATICA, 135
  in OCTAVE, 78–79
integer
  signed, 6
  unsigned, 6
interference patterns, see also diffraction patterns
  Fabry-Perot, 83
  from four slits, 85
internal storage of expressions
  in MATHEMATICA, 95–96
interpolation functions, see shape functions
interpolation points, 277, 278
interpreting source code, see also Local Guide, 152
interrupt
  in MATHEMATICA, 92
  in OCTAVE, 16
ispell, 491–492
iterative algorithm, 373

K

Kaprekar Constant, 90
keywords
  in OCTAVE, see OCTAVE properties
Kirchhoff’s laws, 350
Knuth, Donald, 459
Kronecker delta, 262

L

Lagrange equations
  for 2D harmonic oscillator, 146
  for compound pendulum, 147
Laguerre polynomials, 144
Laguerre’s method, see root finding
Lampert, Leslie, 459
LAPACK, 1
  contact information, 507
Laplace transforms, 13–14, 197–198, 249
  definition, 13, 114, 197
  for ODEs, 197
  in MATHEMATICA, 114, 148
Laplace’s equation, 352, 361, 362
  in 2D, 396–399
  in OCTAVE, 410–416
  in OCTAVE, 167–168
  in pseudocode, 162–164
\LaTeX, 1
  accents in, 495
  counters in, 496
  dashes in, 471
  math mode in, 472–473
  quotation marks in, 472
  special characters in, 496
latex, 462, 463, 481
\LaTeX{} commands
  \textbackslash (italic correction), 468
  \textbackslash (new line), 470, 475
  \textbackslash! (negative thin space), 473
  \textbackslash (thin space), 473
  \textbackslash - (allow hyphen), 472
  \textbackslash: (medium space), 473
  \textbackslash; (thick space), 473
  \addcontentsline, 482–483
  \addtocontents, 482–483
  \Alph, 474, 497
  \alph, 474, 497
  \and, 488
  \arabic, 474
  \author, 492
  \bf, 492
  \bfseries, 468
  \boldmath, 473
  \caption, 475, 476, 478, 481
  \centerline, 478
  \chapter, 472
  \chaptername, 499
  \char, 496
  \clearpage, 469
  \contentsname, 499
  \date, 492
  \definecolor, 485, 489
  \documentclass, 460, 462, 465, 486, 492
  \documentstyle, 460
  \fbox, 496
  \figurename, 498
  \footnote, 472, 492
  \footnotesize, 469
  \frac, 492
  \framebox, 496
  \hspace, 469, 470
  \hspace*, 469
  \Huge, 469
  \huge, 469
  \IfFileExists, 489
  \ifthenelse, 488–489
  \includegraphics, 476–479, 486
  \index, 483
  \indexname, 499
  \input, 495
  \int, 492
  \item, 473–474, 497
  \itshape, 468
  \label, 472, 475, 476, 478, 481, 492
  \labelitem, 474
  \labelitemi, 474
  \labelitemii, 474
  \labelitemiii, 474
  \LARGE, 469
  \Large, 469, 492
\large, 469
\LaTeX, 492
\listoffigures, 482–483
\listoftables, 482–483
\makebox, 496
\makeindex, 483–484
\maketitle, 492
\mathbf, 473
\mathcal, 495
\mathit, 473
\mathrm, 473
\mathtt, 473
\mbox, 496
\mdseries, 468
\Needspace, 499
\needspace, 499
\newboolean, 488–489
\newcommand, 493
\newcounter, 497
\newlength, 470
\newpage, 469
\noindent, 469, 492
\normalfont, 468
\normalsize, 469
\not, 488
\or, 488
\pageref, 472, 475, 476
\pagestyle, 492, 495
\parbox, 496
\part, 472
\partial, 492
\printindex, 483
\quad, 469
\ref, 472, 475, 476, 492
\renewcommand, 471, 474, 494, 497, 498
\rmfamily, 468
\Roman, 474, 498
\roman, 474, 498
\rule, 496
\scriptsize, 469
\scshape, 468
\section, 472
\setboolean, 488–489
\setcounter, 497
\setlength, 465, 469, 492
\settodepth, 469
\settoheight, 469
\settowidth, 470
\sfamily, 468
\slshape, 468
\small, 469
\subsection, 472
\subsubsection, 472
\tablename, 499
\tableofcontents, 482–483
\textbf, 468
\textit, 468
\textmd, 468
\textnormal, 468
\textrem, 468
\textsc, 468
\textsf, 468
\textsl, 468
\texttt, 468
\textup, 468
\the..., 497–498
\theenumi, 474
\theenumii, 474
\theenumiii, 474
\theenumiv, 474
\thispagestyle, 492, 495
\tiny, 469
\title, 492
\today, 493, 494
\ttfamily, 468
\twocolumn, 498
\typein, 489
\typeout, 489
\upshape, 468
\usepackage, 471, 477, 483
\value, 497
\verb, 468, 492, 496
\verbatiminput, 496
\vspace, 469, 476
\vspace*, 469
\LaTeX contact information, 459, 509
\LaTeX counters
enumi, 474, 497
enumii, 474
enumiii, 474
enumiv, 474
figure, 497
footnote, 497
page, 497
secnumdepth, 482
section, 497
subsection, 497
subsubsection, 497
table, 497
tocdepth, 482
\LaTeX declarations, see \LaTeX commands
\LaTeX document classes, 461
article, 465, 472
book, 465, 472
report, 465
slides, 465
\LaTeX environments, 470–471
\(...\) (in-line equation), 472
\[\ldots\] (displayed equation), 472
$\ldots$ (in-line equation), 472
array, 499
center, 470, 475, 477, 481, 492
comment, 496
displaymath, 472
document, 461, 462, 492
enumerate, 473–474, 496, 497
eqnarray, 472
equation, 472, 492
figure, 476–479, 481
flushleft, 492
itemize, 473–474
minipage, 496
pict2e, 482
picture, 476, 482
quotation, 470
quote, 470
table, 474–476
tabular, 474–476
tikzpicture, 479–482
\\baselineskip, 469, 471
\\baselinestretch, 471
\\bottomfraction, 497
\\evensidemargin, 465, 492
\\floatpagefraction, 497
\\itemsep, 474
\\marginskip, 484
\\marginparwidth, 496
\\oddsidemargin, 465, 492
\\oddsidemargin, 465, 492
\\parindent, 465, 492
\\parskip, 465, 492
\\textfraction, 497
\\textwidth, 465, 492
\\topfraction, 497
\\topmargin, 465, 492

LATEX length parameters, 466
LATEX Manual, 459
LATEX output from MATHEMATICA, 132–133
LATEX packages, 471
\\alltt, 495
\\color, 485, 489
\\graphicx, 476–479, 485–486
\\hyperref, 485–486, 489
\\ifthen, 488–489
\\imakeidx, 483–484
\\makeidx, 483–484
\\needspace, 499
\\notebook2e, 133
\\showidx, 484
\\tikz, 476, 479–482
\\verbatim, 495–496

LATEX page previewers, see also xdvi, yap, 490–491
LATEX preamble, 465–467
LATEX sample, 461–464, 492–493
LATEX special symbols
’’ (closing double quotation), 472
--- (em dash), 471
– (en dash), 471
~ (dash), 471
\ (minus sign), 472
\ (hard space), 471
‘‘ (opening double quotation), 472
LATEX templates, 466
LATEX type size, 468
LATEX type styles, 467–468
lawn mower, 345
least squares fitting, 350, 352
Legendre polynomials, 142, 284, 302, 347
Lennard-Jones potential, 349
licensing of software, 14
line labels
in MATHEMATICA, 92
linking object code, see Local Guide, 152
lists
in MATHEMATICA, 94–96, 99–100, 205
Local Guide, 2, 8, 9, 11, 14–16, 35, 37, 91, 92,
125, 135, 137, 168, 221, 244, 266, 284,
289, 323, 459, 461–465, 467, 477, 490,
492
logical conditions, see logical expressions
logical expressions
in LATEX, 488–489
in MATHEMATICA, 123–124
in OCTAVE, 34–35
in pseudocode, 155–156
logical operators, see logical expressions
logical unit number (LUN)
in OCTAVE, 31
logistic growth, 190, 201–202
long integer
signed, 6
unsigned, 6
loops
in MATHEMATICA, 103, 108, 122–123
in OCTAVE, 32–35
in pseudocode, 157–159
Lorentz distribution, 83, 300, 304
Lorentz lineshape, 39
lp, see UNIX commands
LSODE, 1
LSODE contact information, 507
LU decomposition, 341–342, 377

M

m-files, see OCTAVE scripts, OCTAVE functions
machine code, 152
Macintosh OS, 1
Mac\TeX
contact information, 508
magnetic field, 260
in wave guide, 66–67, 88, 170–173
of circular loops, 43, 85, 179, 268–272, 305
of Helmholtz coil, 268–272
of solenoid, 85
magnetic vector potential, 260
MAPLE, 1
MAPLE contact information, 508
MATHEMATICA, 1
mathematica, 92
MATHEMATICA command files, 124–126
crossdot, 125, 150, 266, 269
testmultiply, 125

MATHEMATICA commands, see also

MATHEMATICA functions, 94

Animate, 133
Apart, 104, 140, 148
Append, 137
Black, 127, 128
Boole, 123
Cancel, 118
Clear, 97, 98
CoefficientArrays, 207
CoefficientList, 142, 210
Collect, 117, 142, 200, 206, 211
ComplexExpand, 105–107, 117
Conjugate, 105
ContourPlot, 129
Cross, 126
Curl, 120, 149, 150
D, 112–116, 126, 142, 145, 245, 317
Dashing, 127, 128
Delete, 137
Det, 110, 208
DiagonalMatrix, 100
Display, 131
Div, 120, 149, 150
Dot, 126
dSolve, 115–119, 198–211
Eigensystem, 111
Eigenvalues, 111
Eigenvectors, 111
Element, 208, 273–274
Evaluate, 246
Exit, 102
Export, 131, 132
ExpToTrig, 106, 107, 206
Factor, 103, 117, 140, 320
Remove, 109
FileNames, 137
FilePrint, 126
FindRoot, 271, 334–337, 342, 345
FullForm, 103, 104, 112, 113, 141
Get, 124, 125, 266, 269
Grad, 120, 149, 150
Graphics, 138, 338
Head, 103, 113
HoldForm, 113
IdentityMatrix, 100
Insert, 137
Integrate, 112, 114, 263–275, 296, 298
Inverse, 110
InverseLaplaceTransform, 114, 148
LaplaceTransform, 114
Laplacian, 120, 149, 150
Limit, 114, 201, 204, 264
Line, 338
LinearSolve, 109, 342
ListPlot, 299, 336, 338
Lucross, 125–126, 269
ludot, 125–126, 266, 269
Map, 117, 338
MatrixForm, 100, 110, 146, 208
N, 102, 109, 265, 271, 272, 296, 298, 316, 317, 332, 342
Names, 97
NDSolve, 245–246
NIntegrate, 296–299
Normal, 113, 114, 207, 267, 268, 271, 273, 319, 320
NSolve, 109, 332–339, 342
Numerator, 271
Off, 136
On, 136
Options, 94, 129, 138
ParametricPlot, 128
ParametricPlot3D, 139
Part, 103–105, 141
Plot, 126–128, 131, 133, 138, 202, 204, 246, 270, 275, 318
Plot3D, 129, 133
Prepend, 137
Print, 103, 108, 124–125
Protect, 98
Put, 132
PutAppend, 132
Quit, 93, 101, 102, 105, 107–109, 111, 113, 114, 118, 129, 200, 265, 317, 318
Range, 99, 298, 299
ReleaseHold, 114
Remove, 97, 98
Replace, see /., 104
ReplacePart, 105, 137, 141
Rest, 338
Root, 315
Series, 112, 114, 200, 265, 267, 268, 270, 272, 273, 319, 320
SetDirectory, 136–137
SetOptions, 137
Show, 138, 339
Solve, 108, 109, 117, 199, 208, 210, 211, 264, 271, 315–319, 332, 342, 345
StyleForm, 127, 128, 202, 204
Sum, 275
Table, 99, 100, 210, 211, 246, 298, 299, 336–339
TeXForm, 132
Text, 127, 138
Thickness, 127, 128, 202, 204, 270, 299, 318, 336, 338
Tr, 110
Transpose, 110
TrigExpand, 107, 108
TrigReduce, 107, 108
TrigToExp, 107
Union, 205
Unprotect, 98
ZeroMatrix, 146

MATHEMATICA constants, 94
E (2.71828...), 94
False, 94
I (\(\sqrt{-1}\)), 94, 105, 106, 205
Infinity (\(\infty\)), 94, 112, 264
Pi (\(\pi\)), 94, 98, 99, 102, 129, 208, 298, 336
True, 94, 200, 201

MATHEMATICA contact information, 508

MATHEMATICA control structures
If, 112, 123–124
While, 123

MATHEMATICA domains
Integers, 208, 273

MATHEMATICA functions, see also MATHEMATICA commands, 94–95, 97–98
Abs, 106
ArcTan, 106
Arg, 106
ConditionalExpression, 112, 334
Cos, 95, 97, 105, 107, 126, 133, 139, 206, 266, 269, 273–275, 299, 318, 334, 335
Cosh, 105
Cot, 338
Exp, 106, 112, 114, 126, 201, 298, 318, 335
Factorial, 102
Im, 105, 106, 112
LegendreP, 143
Log, 268
Re, 105, 106
Sin, 98–100, 105, 107, 113, 114, 126, 129, 131, 133, 138, 139, 206, 263, 266, 269, 273–275, 318, 337
Sinh, 105
Sqrt, 102, 110, 117, 205, 206, 264, 266, 268, 272, 298, 299, 317, 338

MATHEMATICA heads
Plus, 95, 96, 103, 104, 113
Power, 96, 103, 104, 113
Rational, 112, 113
SeriesData, 112
Times, 95, 96, 103, 104, 113

MATHEMATICA logical expressions, see logical expressions in MATHEMATICA

MATHEMATICA manuals, 91
MATHEMATICA notebooks, 134–135
MATHEMATICA options, 94
AccuracyGoal, 246, 297, 334
AnimationRate, 133
AspectRatio, 128, 129
Assumptions, 94, 112, 265, 267, 268
AxesLabel, 127, 128, 131, 202, 204, 246, 270, 275, 318, 336, 338
AxesOrigin, 338

Cartesian, 119–121
ColorFunction, 129
Contours, 129
ContourShading, 129
Cylindrical, 119–121
DisplayFunction, 139
Epilog, 127
FontFamily, 127, 128, 202, 204
FontSize, 127, 128, 202, 204
FontWeight, 127, 128, 202, 204
FormatType, 131
GaussPoints, 297
GridLines, 299, 338
MaxIterations, 334
MaxRecursion, 297
MaxSteps, 246
MaxStepSize, 246
Method, 297, 298
PlotJoined, 299, 336, 338
PlotLabel, 127, 128, 202, 204, 299
PlotPoints, 127–129
PlotRange, 94, 127, 128, 133, 202, 204, 270, 275, 299, 318, 338
PlotStyle, 127, 128, 139, 202, 204, 246, 270, 275, 299, 318, 336, 338
PrecisionGoal, 246, 297
Spherical, 119–121
TargetFunctions, 94, 106–107
TextStyle, 94
Thickness, 139
WorkingPrecision, 246, 297, 334

MATHEMATICA packages, 121–122
LinearAlgebra

MATHEMATICA special symbols
! (factorial), 102
'(differentiation), 115, 118, 199–202, 205, 206, 210
(*...*) (comments), 125
* (multiplication), 94, 95, 97, 102, 110
+ (addition), 94, 95, 101
- (subtraction), 94, 95, 102
. (clear variable), 98
. (matrix multiplication), 110
/ (division), 94, 95, 101
// (convert form), 100, 102, 271, 316
:= (function definition), 94, 97–98, 113, 274, 338
e (assignment), 97, 99
== (equality), 94, 108, 109, 112, 115, 117, 118, 199, 202, 205–207, 211, 245, 264,
271, 315–317, 334, 335
? (help), 93, 97, 98
?? (more help), 93
[[...]] (element of list), 99, 100, 108, 116, 117, 199, 201, 204, 208, 210, 211, 264, 298, 318, 319, 337, 338
\#1 (generic variable), 315
% (previous output), 100, 103, 107–110, 112, 114, 211, 267, 271, 273, 317, 319, 335
%% (second previous output), 103, 112
%%% (third previous output), 103
\%n (output on line \(n\)), 96, 103, 211
\.
(pattern), 97–98, 113, 274, 298, 299
>> (direct output to file), 132
>>> (append output to file), 132
^ (exponentiation), 95, 97, 98, 102, 103
Out[n] (output on line \(n\)), 96, 103

MATHEMATICA statements
termination of, 92

MATHEMATICA system variables, 94
$BaseDirectory, 135
$HomeDirectory, 94
$ignoreEOF, 94
$MachinePrecision, 246, 297
$MessageGroups, 136
$packages, 94, 121
$Path, 94, 137
$RecursionLimit, 94
$TextStyle, 94
$TopDirectory, 94
$UserBaseDirectory, 135

mathematical functions
in MATHEMATICA, see MATHEMATICA functions
in OCTAVE, see also OCTAVE functions,
OCTAVE programs, OCTAVE commands, 22

MATLAB, 1, 15
MATLAB contact information, 508
MATLAB system variables
ans, 18
matrices
in MATHEMATICA, 99, 100, 110–111
in OCTAVE, 19–24
matrix multiplication
in MATHEMATICA, 110
in OCTAVE, 23–24
max-min problems, 308
MAXIMA, 1
MAXIMA contact information, 508
Maxwell-Boltzmann distribution, 303
method of bisection, see root finding
method of Frobenius, 209–211
method of least squares, see least squares fitting
midpoint method, 217
midpoint rule, see quadrature
MiKTeX, 1
MiKTeX contact information, 508
ML (left mouse button), 9
MM (middle mouse button), 9
moment of inertia
definition of, 257
of cardioid, 302
of semicircular plate, 257, 266, 290–291
motion
of membrane, 69–70, 89
of string, 89
MR (right mouse button), 9
MUDPACK, 1
contact information, 508
Müller’s method, see root finding
multi-way selection
in pseudocode, 157
multistep methods
for ODEs, 220–221

N

NAG library contact information, 508
named common
in OCTAVE, see global variables
natural frequencies
of spherical cavity, 348
of stiff bar, 347
of three-mass system, 346
of two-mass system, 309–310, 317–318, 326–327
Navier-Stokes’ equations, 363
Newton’s method, see root finding
Newton’s second law, 255, 256, 309, 311
in fluids, 365–367
Newtonian fluid, 307
node numbers
global, 378
local, 378
nodes, 355
non-viscous fluid flow, 367
normal distribution, see Gaussian distribution
normal modes
of square membrane, 51–58
of string, 89
normal stress
in fluids, 365–367
numerical approximation
for first derivative, 370
for second derivative, 371
numerical integration, see quadrature
Numerical Recipes, 1
NUMERICAL RECIPES contact information, 509
nybble, 5
### O

- object code, 152
- object-image relations, 84
- object-oriented programming, 28
- octal number system, 5
- OCTAVE, 1
  - current directory, 50
- OCTAVE command files, see OCTAVE scripts, OCTAVE programs
- OCTAVE commands, see also OCTAVE functions, OCTAVE programs
  - `addpath`, 78
  - `and`, 34
  - `axes`, 29, 71
  - `ceil`, 17
  - `floor`, 17
  - `clabel`, 58
  - `clear`, 20, 33, 34, 77, 286
  - `clf`, 74
  - `close`, 64, 74
  - `colorbar`, 63, 64
  - `colormap`, 55, 57, 64
  - `compan`, 323, 325
  - `contour`, 51, 57–59, 61, 67
  - `cross`, 24
  - `cumsun`, 81
  - `det`, 23
  - `diary`, 32, 49
  - `disp`, 80, 399
  - `dot`, 23
  - `double`, 17
  - `errorbar`, 48
  - `exit`, 16
  - `eye`, 20
  - `fclose`, 31, 35, 38, 47, 69
  - `fgetl`, 35, 38
  - `figure`, 29, 74–75
  - `fix`, 17, 22, 401, 412, 414, 428
  - `fopen`, 31, 35, 38, 47, 69
  - for matrix multiplication, 23–24
  - `format`, 21, 79, 284, 286–289, 325, 326, 329
  - `fprintf`, 31, 80
  - `fscanf`, 31, 35, 38, 47, 69
  - `fzero`, 324–326, 329
  - `set`, 67
  - `gcf`, 30, 74–75
  - general structure of, 18–19
  - `genpath`, 78
  - `get`, 29, 30, 63, 72, 74–75
  - `getframe`, 69, 70
  - `global`, 40, 227, 229, 231, 234, 236, 238, 243, 328, 329
  - `graphics_toolkit`, 26, 35, 38, 43, 48, 52, 59, 61, 222, 292, 296, 327–329, 332
  - `grid`, 26, 27, 29, 73, 234, 236, 292, 294, 296, 327–329, 375, 389
  - `quadl`, 294–295
  - `help`, 37, 41, 244
  - `hold`, 35, 36, 38, 45, 48, 58, 64, 67, 76, 222, 227, 229, 234, 239, 241, 327, 328, 332, 375
  - `input`, 79, 373, 386, 399, 401, 404, 407, 412, 415, 428
  - `int16`, 17
  - `int2str`, 17
  - `int32`, 17
  - `int64`, 17
  - `int8`, 17
  - `inv`, 23, 24
  - `isosurface`, 64–65
  - `length`, 296
  - `line`, 29
  - `linspace`, 25, 288, 291–293
  - `load`, 30–31
  - `loglog`, 48
  - `logspace`, 49
  - `lookfor`, 41
  - `lsode_options`, 242–243
  - `lu`, 81
  - `lubisect`, 326, 331
  - `magic`, 20
  - `max`, 65, 329, 390, 402–404, 410
  - `mesh`, 51–54, 59, 61, 69, 70, 167, 415, 435
  - `meshc`, 59
  - `meshgrid`, 51–52, 59, 61, 63, 67, 70, 168, 170, 172, 173, 435
  - `min`, 65, 390, 402–404
  - `movie`, 69–70
  - `nor`, 34
  - `not`, 34
  - `num2str`, 17, 80
  - `or`, 34
  - `patch`, 64–65
  - `path`, 78
  - `pause`, 401, 407, 415
  - `plot3`, 73
  - `plotradio`, 37
  - `plotrc`, 49
  - `pol2cart`, 51, 59
  - `polar`, 46
  - `print`, 50
  - `printf`, 412, 414

---

**Note:** The above text is an excerpt from a document, likely a manual or a technical reference, listing a variety of commands and functions in the OCTAVE programming language. Each entry is labeled with its function or command name, followed by a brief description or usage context. The entries are organized alphabetically, providing a comprehensive list for users of the OCTAVE environment.
quad, 293
quad2d, 288–291
quad2v, 288–291, 294–296
quit, 16
quit3, 69, 88
rand, 20
replacedefaultpath, 78
return, 399, 405
rmpath, 78
roots, 323, 325, 327
rot90, 435, 436
rotate, 63
round, 17
save, 30–31
semilogx, 47–49
semilogy, 48
zeros, 387
set, 26, 27, 29, 30, 43, 46, 48, 59, 61–65, 67,
71, 72, 74–75, 222, 227, 234, 236, 239,
241, 292, 294, 296, 327–329, 332,
400–401, 404, 406–407, 409–410, 435
shading, 55–57, 62–64
single, 17
size, 239, 241
slice, 62–64
sort, 81
sprintf, 222
str2num, 17
subplot, 46–47, 76, 229, 375, 389, 404,
409–410
sum, 81, 284, 287
surf, 51, 55–57, 59, 61, 63, 168
surfc, 59
surf1, 59
system, 16
text, 35, 36, 38, 47, 239, 241, 332
title, 26, 27, 29, 43, 45, 46, 48, 53, 57, 73,
76, 222, 227, 229, 292, 294, 296, 375,
389, 404, 409–410
toascii, 17
trace, 23
transpose, 23, 25, 286, 326, 327
trapz, 287–288, 291
uint16, 17
uint32, 17
uint64, 17
uint8, 17
view, 54, 64–65, 416, 435
who, 20, 37
whos, 20, 37
xlabel, 26, 27, 29, 43, 48, 53, 57, 67, 73,
222, 227, 229, 236, 239, 241, 292, 294,
296, 327, 329, 435
ylabel, 26, 27, 29, 43, 48, 53, 57, 67, 73,
222, 227, 229, 236, 239, 241, 292, 294,
296, 327, 329, 435
zeros, 20, 33, 63, 167, 406, 410, 411,
430–432, 434
zlabel, 53, 57
OCTAVE contact information, 509
OCTAVE control structures
break, 80
for/end, 32–34, 61, 69, 81, 167, 221, 222,
234, 236, 239, 241, 284, 286, 287, 292,
293, 296, 326–329, 332, 373, 376,
386–387, 390, 400, 406–407, 410, 412,
428–430, 432–434
function/endfunction, 38–40, 224–226,
228, 229, 231, 234, 238, 288, 290, 293,
295, 296, 325, 328, 331
if/elseif/elseif/end, 34, 399, 401, 405,
430, 432–433
return, 80–81
while/end, 34
OCTAVE functions, see also OCTAVE
commands, OCTAVE programs, 37–41
abs, 35
apollo, 244
available_graphics_toolkits, 26
built-in, 22
cos, 44, 46, 47, 59, 67, 70, 73, 75, 76, 172,
293, 296, 329, 331
sin, 168
cosh, 26, 47
dampharm, 244
damposc, 229, 244
decay, 226, 244
decaymod, 228, 244
drvnosc, 244
drvnpend, 244
eig, 41–42
exp, 45, 49, 59, 61, 170, 284, 286–288, 328,
329
fix, 399, 400, 404, 406–407
fsolve, 345
gausint, 288, 292
henon, 244
largamp, 244
lineshape, 39–40
lorenz, 244
lubisect, 324–325, 354
lugen_list, 168–170, 173, 174
luplot, 37–41
luread_list, 174
max, 376
min, 376
moment, 290
onedshm, 244
planet, 231, 244
projectile, 328–329
qmshm, 238, 244
qmupper, 331
rings, 296
rossler, 244
sin, 22, 44, 45, 47, 51, 52, 67, 70, 73, 75, 76,
168, 172, 293, 328, 331
sinh, 47
INDEX

spiralc, 293
spirals, 293
sqrt, 22, 49, 61, 170, 284, 286–288, 290, 291, 296, 327, 331
stdwaves, 234, 244
turnpts, 325–326
twoshm, 244
twooscil, 245
user-defined, 37
vandpol, 245
zeros, 373, 400

OCTAVE m-files, see OCTAVE scripts, OCTAVE functions
OCTAVE objects, 28–29
OCTAVE programs, see also OCTAVE commands, OCTAVE functions
fdm1d.m, 374, 445
fdmdiffus1d.m, 407, 449–450
fdmlap2d.m, 412, 450–451
fdmlap2d_plot.m, 415, 451–452
fdmwave1d.m, 401–402, 448–449
fem1d.m, 387, 446–447
fem2d.m, 434, 456–458
lugen_list.m, 181–182
luread_list.m, 183

OCTAVE properties, 18–19, 28–29
edgecolor, 53, 57, 59, 61, 65, 69, 415, 435
facecolor, 65
fontangle, 71
fontname, 71–72
fontunits, 28, 71
fontweight, 71
interpreter, 71
linestyle, 222, 227, 229, 235
marker, 43–44
markerfacecolor, 44
markersize, 44
menubar, 74
name, 74, 75
position, 74, 75
tolx, 324
visible, 62, 63, 332
xdata, 63
xgrid, 30
xlim, 29, 30, 234, 328, 332
ydata, 63
ygrid, 30
ylim, 29, 30, 234, 327, 328, 332, 400–401, 404, 406–407, 409–410
zdata, 63

OCTAVE special symbols, 73
! (logical not), 34
\n (new line), 28, 31, 80, 222, 412, 414
\ldots (continue statement on next line), 16, 29, 30, 61, 63, 71–74, 76
\' (transposition), 23, 35, 47, 69, 435
\ldots' (enclose strings), 18, 19
* (multiple character wildcard), 77
* (multiplication), 22–23, 32
+ (addition), 22
- (subtraction), 22
.* (multiplication), 22, 32, 45, 51, 52, 59, 61, 67, 70, 168, 170, 172, 173, 290, 291, 331
./ (division), 22, 23, 39, 40, 49, 51, 52, 168, 331
.\ldots (exponentiation), 22, 43, 49, 52, 59, 61, 73, 88, 168, 170, 173, 284, 286, 288, 290, 291
/ (division), 22, 25, 343
; (range), 21, 25, 51–52, 61, 172, 173, 326–329, 332
; (separate rows in array), 20
; (separate elements in column vector), 227, 230, 234, 238
; (suppress output), 16, 20, 61, 63, 221, 226–228, 230, 234, 238, 284, 286–288, 292–294, 326–330, 332
= (assignment), 18
== (equality), 401, 407, 414
? (single character wildcard), 77
# (comment), 16, 226, 228
% (comment), 16
%d (format), 31, 412, 414
%e (format), 31
%f (format), 31, 47, 222
\ldivision)24, 343, 374, 387, 434
\ldots' (enclose strings), 18
\ldash (exponentiation), 22, 39, 40, 331
\ldash (logical not), 34
\ldash (logical not equal), 433
for expressing conditions, 34

OCTAVE symbol codes, 44
OCTAVE system variables
I (\sqrt{−1}), 19
i (\sqrt{−1}), 19
J (\sqrt{−1}), 19
j (\sqrt{−1}), 19
pi (\pi), 22, 49, 51, 52, 61, 67, 70, 76, 168, 170, 172, 284, 286–288, 290, 291, 328,
INDEX

329
on-line help
in MATHEMATICA, 93
in OCTAVE, 37
ordinary differential equations
with MATHEMATICA, 115–119, 198–211, 245–246
with OCTAVE, 221–245
orthogonal functions, 262–263
OzTeX contact information, 509

P
page previewer
in LATEX, 490–491
partial fractions, 104, 148, 248
passwords, see Local Guide
path (in directory tree), 8
path specification
in MATHEMATICA, 136
in OCTAVE, 77
PDEs
classification of, 362–363
PDF files, 463–464
dfcrop, 486–487
dflatex, 463, 480, 485–486
dftops, 487
pendulum, simple, see simple pendulum
pivotal condensation, 341
pivoting, 341
placing text in graphics
in MATHEMATICA, see MATHEMATICA commands, Text; MATHEMATICA options, Epilog
in OCTAVE, see OCTAVE commands, text
Planck radiation law, 85
Poisson’s equation, 362
polar coordinates
in MATHEMATICA, 106–107, 119–121
in OCTAVE, 45–46, 59
population biology, 190, 201–202, 253
preallocation of arrays
in OCTAVE, 33
predictor-corrector methods, 215
procedure, 160
program, 151
projectile motion, see trajectories
properties
in OCTAVE, see OCTAVE properties
proprietary software, 14
protected variables
in MATHEMATICA, 98–99
in OCTAVE, see OCTAVE special symbols, OCTAVE system variables
ps2pdf, 464, 481, 486–487
pseudocode comments, 157, 160
pseudocode control structures, 151, 156–160
CASE-END_CASE, 157
EXECUTE, 159
FOR-THRU-DO, 158–159, 164–166
FOR-THRU-DO-STEP, 158
IF-THEN-ELSE-END_IF, 3, 4, 157, 161, 162
LOOP-END_LOOP, 3, 4, 157, 160–162
REPEAT-UNTIL, 158
WHILE-DO, 158
pseudocode directives
DIMENSION, 153, 164, 166
pseudocommands
ATTACH, 166
CLOSE, 166
READ, 154
WRITE, 154, 164, 166
PYTHON, 1
PYTHON scripts
ExtractFileName.py, 506
Q
Q of resonant circuit, 86
quadratic formula, 108, 314, 315, 345, 347
quadrature, 276–284
adaptive algorithms, 281
assessing error in, 279–281, 284–287
Gauss-Kronrod, 298
Gauss-Legendre, 284
Gaussian, 281–284, 286–287, 303
iterative algorithms, 281
midpoint rule, 148, 276
Newton-Cotes, 276–281
recursive trapezoidal rule, 278
Romberg integration, 279
roundoff error, see roundoff error
Simpson’s rule, 148, 277, 279, 283, 286, 303
trapezoidal rule, 276, 284
truncation error, see truncation error
with MATHEMATICA, 296–299
with OCTAVE, 284–296
quality factor, see Q of resonant circuit
quantum energy levels, see energy levels
quantum probabilities, 262
in hydrogen atom, 60, 86, 169–170, 179, 301, 305
of harmonic oscillator, 262, 291, 298, 303
quantum scattering, 143
quantum wave function, see wave function
R
radius of gyration
definition of, 257
of cardioid, 302
of semicircular plate, 266
railroad problem, 348
range of projectiles, see trajectories
Rayleigh criterion, 349
reaction kinetics, 253
reading data
from files
  in OCTAVE, 35–36, 174
  in pseudocode, 167
from keyboard
  in OCTAVE, 79
  in pseudocode, 154
recasting
  algorithms, 175
  control structures, 175
recursion relationship, 209
reflection at dielectric film, 84
relativity
  mass/momentum/energy, 84
  motion under constant force, 264–265
relaxation, 163–164
residual, 323
resonance
  in LRC circuit, 86, 192, 249
restoring initial state
  in MATHEMATICA, 101
  in OCTAVE, 79
returning to command level
  in MATHEMATICA, 92
  in OCTAVE, 16
returning to operating system
  from MATHEMATICA, 93
  from OCTAVE, 16
ReVTeX, 495
Richardson extrapolation, 278, 285
Ritz variational method, 377
Rodriques’ formula, 142
Romberg integration, see quadrature
root finding
  accuracy in, 322–323
  by bisection, 320, 324–325
  by Laguerre’s method, 322
  by Miller’s method, 322
  by Newton’s method, 321, 334, 343–345, 347
  by secant method, 322, 334
  by trial and error, 343
  in MATHEMATICA, 315–320, 332–339, 345
  in OCTAVE, 323–332, 342–343, 345
roundoff error, 372–373, 377, 399
  in numerical quadrature, 279
  in ODEs, 218
rubber lengths in \LaTeX, 465
Ruchardt’s experiment, 352
Runge-Kutta methods, 216–218
running program, see Local Guide

S

saving files, see writing data to files
scalar fields
  three-dimensional, 169–170
    in OCTAVE, 170
  two-dimensional, 168–169
    in OCTAVE, 168–169
search path
  in MATHEMATICA, 137
  in OCTAVE, 35, 78
secant method, see root finding
self-adjoint ODE, 369
separating statements on one line
  in MATHEMATICA, 92
  in OCTAVE, 16, 33, 34
  in pseudocode, 156
series solutions to ODEs, 209–211
shape functions, 377–379
shear stress
  in fluids, 365–367
shooting method, 232
sigmoid curve, 190, 203
simple pendulum, 251, 258, 272–273, 305
Simpson’s rule, see quadrature
simultaneous equations
  by direct methods, 342
  by iterative methods, 342
  linear, 340–343
  nonlinear, 343–345
singular equations, 340
singular value decomposition, 342
solving Laplace’s equation, see Laplace’s equation
solving linear equations, 340–343
  in MATHEMATICA, 109
  in OCTAVE, 24–25
solving ODEs
  with MATHEMATICA, 198–211, 245–246
  with OCTAVE, 221–245
sound waves in fluids, 368–369
source code, 152
spawning to operating system
  from OCTAVE, 16
special symbols
  in MATHEMATICA, see MATHEMATICA special symbols
  in OCTAVE, see OCTAVE special symbols
speed of sound, 368
spherical coordinates
  in MATHEMATICA, 150
spurious roots, 307, 313, 332, 338
stable equilibrium, 308
standing waves
  in cube, 87, 180
Stark effect, 82, 146, 346
startup files, see initialization files
structure of data files, 164–165
subroutines
  in MATHEMATICA, see MATHEMATICA functions
  in OCTAVE, see OCTAVE functions, OCTAVE programs
suppressing evaluation
in MATHEMATICA, 113–114
suppressing output
  in MATHEMATICA, 92
  in OCTAVE, 16, 20

T

tables
  in L\LaTeX, 474–476
Taylor series
  for solving ODEs, 216–219
  for solving simultaneous equations, 344
  in MATHEMATICA, 112, 142, 148
termination of statements
  in MATHEMATICA, 92
  in OCTAVE, 16
  in pseudocode, 156
\LaTeX, 459, 467
\LaTeX contact information, 459, 509
\LaTeX output
  from MATHEMATICA, see \LaTeX output
texdoc, 479, 500
\TeXlive, 1
\TeXlive contact information, 509
text editors, see Local Guide
text in graphics, see placing text in graphics
TGIF, 1
TGIF contact information, 510
thermal equilibrium, 86
thin lens equation, 84
tikz, 479–482
tikz commands
  \definecolor, 480
  \draw, 480–481
  \foreach, 480, 481
  \node, 480, 481
  \x, 480, 481
trajectories, 255
double pendulum, 248
earth falling into sun, 301, 304
exponentially decaying force, 300
in central forces, 193–195, 230–232, 251, 252
in constant magnetic field, 73, 138, 179
in Coulomb field, 249
in crossed E and B, 252
simple harmonic oscillator, 300
under constant force, 264–265, 300
transformation rules
  in MATHEMATICA, 104, 108, 198, 208
  transmission at dielectric film, 84
transpose
  in MATHEMATICA, 110, 125
  in OCTAVE, 23
  trapezoidal rule, see quadrature
tree-structured directories, 8
tridiagonal systems, 342, 352

truncation error
  global, 219
  in midpoint formula, 280
  in numerical quadrature, 279
  in ODEs, 218, 222–224
  in Simpson’s rule, 280
  in trapezoidal rule, 280, 285
    per step, 219
truth table, 155
turning points, 262, 291, 303, 308, 316, 325–326, 335, 348, 349
two’s complement notation, 6
two-way selection
  in MATHEMATICA, 123–124
  in OCTAVE, 34–35
  in pseudocode, 157

U

UNIX, 1
UNIX commands
  aspell, 491–492
  cleanbak, 500
  cleanTEX, 500
  dvipdfm, 464
  dvips, 463
  ispell, 491–492
  latex, 462, 463
  lp, 464
  ls, 487
  pdfcrop, 486–487
  pdflatex, 463
  pdftops, 487
  ps2pdf, 464
  texdoc, 479
  xdvi, 464
UNIX shell scripts
  ceps2pdf, 487, 505
  cleanbak, 500
  cleanTEX, 500
  ceps2pdf, 487, 505
  rdfileeps, 487, 505
  rdfileps, 487, 505
UNIX tutorial, see Local Guide
unstable equilibrium, 308
user accounts, see Local Guide
user-defined commands, see command files
user-defined functions
  in MATHEMATICA, see MATHEMATICA functions
  in OCTAVE, see OCTAVE functions

V

Van der Pol oscillator, 251
variable names
  in MATHEMATICA, 94
  in OCTAVE, 17
  in pseudocode, 152
INDEX

vector arrays
  three-dimensional, 173–174
  two-dimensional, 170–173
vector calculus
  with MATHEMATICA, 119–121
vector fields
  three-dimensional
    in OCTAVE, 173–174
  two-dimensional
    in OCTAVE, 172–173
vector identities, 150
velocity field
  of rotating object, 180
vertices, 355
vim, 1
viscosity, 367
viscous damping, 187, 190
W

  wave equation
    in 1D, 357–359
    in OCTAVE, 399–404
    in 2D, 391–393
    in 3D, 359–360
wave function, 196–197, 237–241, 262, 302, 303
weighting functions, 380
Wheatstone bridge, 143
Windows, 1
Windows batch files
  ceps2pdf.bat, 487, 504
cleanbak.bat, 500
cleanTEX.bat, 500
cps2pdf.bat, 487, 504
rdfileps.bat, 487, 504
Windows commands
cleanbak, 500
cleanTEX, 500
dir, 487
dvipdfm, 464
dvips, 463
latex, 462, 463
pdfcrop, 486–487
pdflatex, 463
pdftops, 487
print, 464
ps2pdf, 464
texdoc, 479
yap, 464
Winedt, 1
Winedt contact information, 510
writing data to files, 179, 180
  in OCTAVE, 31, 168–170, 172–174
  in pseudocode, 164–167
writing data to screen
  in OCTAVE, 19–20, 80
  in pseudocode, 154
X

  xdvi, 464, 490–491
Xemacs, 1
Xemacs contact information, 510
Xv contact information, 510
Y

  yap, 464, 490–491