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 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.

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

---

1 The first edition was published in 2003, though it experienced a number of edits and adjustments in subsequent years.
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 ifthen and makeidx 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. Gehlbe, 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 Garbachik, 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 members¹ 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, 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), Kao-Biu Luk (University of California, Berkeley), Daryl Macon (Boise State University), Viktor Martisevita (Central College), Donald Miller (Central Missouri State University), Dorn Peterson (James Madison University), Robert Ragan (University of Wisconsin – Lacrosse), Shafiq Rahman (Allegheny College), Rahmatullah Syed (Norwich University), Jorge Talamanes (California State University – Bakersfield), Mark Timko (Elmhurst College), Paul Tjossem

vii
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.
- Peihong 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;

(Grinnell College), and John Walkup (California Polytechnic Institute – San Luis Obispo). In July, 2002 (second workshop): Julio Blanco (California State University—Northridge), William Briscoe (George Washington University), Paul Bunson (Lawrence University), Albert Chen (Oklahoma Baptist University), K. Kelvin Cheng (Texas Tech University), Bob Davis (Taylor University), Mirela Petca (University of Richmond), David Grumbine (St. Vincent College), Kristi Hendrickson BU ’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 routines 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. Any 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.

3Specifically flmoon, xflmoon, caldat, julday, xjulday, aequal, xavevar, x_four, xrk4, xrk4, rkqs, rkck, mrid, hestep, rkdumb, odeint, trapzd, xtrapzd, xtrap, qtrap, xqtrap, qsimp, qromb, point, rtbis, rttris, rtrnewt, rtrnewt, rtsafe, xtsafe, zbrak, gausaj, ludecmp, lubksb, tridiag, svdcmp, svbksb, mnewt, newt, and brydyn (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. 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

1The 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 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.
• 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

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

  ○ 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.\(^2\)
  ○ 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.\(^3\)

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 MiKT\LaTeX\ 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 \emph{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  
1.1 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  

3 Introduction to MATLAB 15  
3.1 Beginning a MATLAB Session 15  
3.2 Basic Entities in MATLAB 17  
3.2.1 Data Types 17  
3.2.2 Variable Names 17  
3.2.3 Assignment of Values to Variables 17  
3.2.4 Commands 18  
3.3 A Sampling of MATLAB Capabilities 19  
3.3.1 Creating and Examining Arrays 19  
3.3.2 Simple Mathematical Manipulations 22  
3.3.3 Matrix Manipulations 23  
3.3.4 Solving Linear Equations 24  
3.3.5 A First Graph 25  
3.4 Properties, Objects, and Handles 28  
3.5 Saving and Retrieving the MATLAB Session 30  
3.6 Loops, Logical Expressions, and Conditionals 32  
3.7 Reading Data from a File 34  
3.8 On-line Help 36  
3.9 M-Files 37  
3.10 Eigenvalues and Eigenvectors 41  
3.11 Graphing Scalar Functions of One Variable 42  
3.11.1 The Basic Strategy 43  
3.11.2 Plotting Several Graphs on One Set of Axes 43  
3.11.3 Polar Plots and Custom Axes 45  
3.11.4 Multiple Plots On a Page 46
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.11.5 Plotting Experimental Data</td>
<td>46</td>
</tr>
<tr>
<td>3.12 Making Hard Copy</td>
<td>49</td>
</tr>
<tr>
<td>3.12.1 ... of Text</td>
<td>49</td>
</tr>
<tr>
<td>3.12.2 ... of Graphs</td>
<td>50</td>
</tr>
<tr>
<td>3.13 Graphing Scalar Functions of Two Variables</td>
<td>50</td>
</tr>
<tr>
<td>3.13.1 A Preliminary: The Function <code>meshgrid</code></td>
<td>51</td>
</tr>
<tr>
<td>3.13.2 Surface Plots: The Function <code>mesh</code></td>
<td>52</td>
</tr>
<tr>
<td>3.13.3 Contour Plots: The Function <code>contour</code></td>
<td>54</td>
</tr>
<tr>
<td>3.13.4 Shaded Surfaces: The Function <code>surf</code></td>
<td>55</td>
</tr>
<tr>
<td>3.13.5 Features of the <code>Figure</code> Window</td>
<td>58</td>
</tr>
<tr>
<td>3.13.6 The Functions <code>meshc</code>, <code>surf</code>, and <code>surfl</code></td>
<td>58</td>
</tr>
<tr>
<td>3.13.7 Functions of Two Variables in Polar Coordinates</td>
<td>59</td>
</tr>
<tr>
<td>3.14 Graphing Scalar Functions of Three Variables</td>
<td>61</td>
</tr>
<tr>
<td>3.14.1 Reduction to Two-Dimensional Displays</td>
<td>61</td>
</tr>
<tr>
<td>3.14.2 Isosurfaces</td>
<td>64</td>
</tr>
<tr>
<td>3.15 Graphing Vector Fields</td>
<td>65</td>
</tr>
<tr>
<td>3.15.1 The Function <code>quiver</code></td>
<td>67</td>
</tr>
<tr>
<td>3.15.2 More Elaborate Displays</td>
<td>68</td>
</tr>
<tr>
<td>3.15.3 Three-Dimensional Vector Fields</td>
<td>68</td>
</tr>
<tr>
<td>3.16 Animation</td>
<td>69</td>
</tr>
<tr>
<td>3.17 Advanced Graphing Features</td>
<td>71</td>
</tr>
<tr>
<td>3.17.1 Fonts</td>
<td>71</td>
</tr>
<tr>
<td>3.17.2 Space Curves</td>
<td>73</td>
</tr>
<tr>
<td>3.17.3 Using Multiple Windows</td>
<td>74</td>
</tr>
<tr>
<td>3.17.4 Customizing Axes</td>
<td>75</td>
</tr>
<tr>
<td>3.17.5 Working with Color</td>
<td>76</td>
</tr>
<tr>
<td>3.18 Miscellaneous Occasionally Useful Tidbits</td>
<td>77</td>
</tr>
<tr>
<td>3.18.1 Specifying Directories</td>
<td>77</td>
</tr>
<tr>
<td>3.18.2 MATLAB's Current Directory</td>
<td>78</td>
</tr>
<tr>
<td>3.18.3 MATLAB's Search Path</td>
<td>78</td>
</tr>
<tr>
<td>3.18.4 Customizing MATLAB</td>
<td>79</td>
</tr>
<tr>
<td>3.18.5 Restoring MATLAB's Initial State</td>
<td>79</td>
</tr>
<tr>
<td>3.18.6 The Command <code>format</code></td>
<td>79</td>
</tr>
<tr>
<td>3.18.7 Suppressing Axes</td>
<td>79</td>
</tr>
<tr>
<td>3.18.8 Positioning Axes</td>
<td>80</td>
</tr>
<tr>
<td>3.18.9 Distinguishing Graphs on a Single Set of Axes</td>
<td>80</td>
</tr>
<tr>
<td>3.18.10 Reading Data from the Keyboard</td>
<td>81</td>
</tr>
<tr>
<td>3.18.11 Writing Data to the Screen</td>
<td>81</td>
</tr>
<tr>
<td>3.18.12 Exiting from Procedures</td>
<td>82</td>
</tr>
<tr>
<td>3.19 References</td>
<td>82</td>
</tr>
<tr>
<td>3.20 Exercises</td>
<td>83</td>
</tr>
<tr>
<td>3.20.1 Writing MATLAB Statements</td>
<td>83</td>
</tr>
<tr>
<td>3.20.2 Finding Eigenvalues and Eigenvectors</td>
<td>84</td>
</tr>
<tr>
<td>3.20.3 Graphing Scalar Functions of a Single Variable</td>
<td>85</td>
</tr>
<tr>
<td>3.20.4 Graphing Scalar Functions of Two Variables</td>
<td>87</td>
</tr>
<tr>
<td>3.20.5 Graphing Scalar Functions of Three Variables</td>
<td>88</td>
</tr>
<tr>
<td>3.20.6 Graphing Vector Fields</td>
<td>89</td>
</tr>
<tr>
<td>3.20.7 Animation</td>
<td>91</td>
</tr>
<tr>
<td>3.20.8 Supplementary Exercises</td>
<td>91</td>
</tr>
</tbody>
</table>

6 Introduction to MAXIMA

6.1 Beginning a MAXIMA Session                                            | 94   |
6.2 On-Line Help                                                          | 95   |
6.3 Basic Entities in MAXIMA                                             | 97   |
6.4 Variable Names in MAXIMA                                             | 97   |
6.5 Expressions in MAXIMA                                                | 97   |
6.6 Binding Expressions to Variables; Defining Functions                 | 99   |
6.7 Using MAXIMA                                                         | 100  |
6.7.1 ... for Arithmetic                                                | 101  |
6.7.2 ... for Algebra                                                   | 102  |
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.7.3</td>
<td>for Complex Variables</td>
<td>104</td>
</tr>
<tr>
<td>6.7.4</td>
<td>for Trigonometry</td>
<td>105</td>
</tr>
<tr>
<td>6.7.5</td>
<td>for Algebraic Equations</td>
<td>106</td>
</tr>
<tr>
<td>6.7.6</td>
<td>for Basic Calculus</td>
<td>108</td>
</tr>
<tr>
<td>6.7.7</td>
<td>for Laplace Transforms</td>
<td>109</td>
</tr>
<tr>
<td>6.7.8</td>
<td>for Ordinary Differential Equations</td>
<td>110</td>
</tr>
<tr>
<td>6.7.9</td>
<td>for Lists, Matrices, and Arrays</td>
<td>113</td>
</tr>
<tr>
<td>6.7.10</td>
<td>for Vector Calculus</td>
<td>115</td>
</tr>
<tr>
<td>6.8</td>
<td>Loops, Logical Expressions, and Conditionals</td>
<td>117</td>
</tr>
<tr>
<td>6.9</td>
<td>Command Files</td>
<td>119</td>
</tr>
<tr>
<td>6.10</td>
<td>High-Resolution Plotting</td>
<td>123</td>
</tr>
<tr>
<td>6.11.1</td>
<td>... of Text</td>
<td>130</td>
</tr>
<tr>
<td>6.11.2</td>
<td>... of Graphs</td>
<td>132</td>
</tr>
<tr>
<td>6.12</td>
<td>Output in ( \text{E}\text{T}\text{X} ) Format</td>
<td>132</td>
</tr>
<tr>
<td>6.13</td>
<td>Animation</td>
<td>133</td>
</tr>
<tr>
<td>6.14</td>
<td>Using the Workbook</td>
<td>133</td>
</tr>
<tr>
<td>6.15.1</td>
<td>Search and Replace: Small Angle Approximations</td>
<td>134</td>
</tr>
<tr>
<td>6.15.2</td>
<td>Simplification Rules</td>
<td>136</td>
</tr>
<tr>
<td>6.15.3</td>
<td>Pattern Matching Functions</td>
<td>137</td>
</tr>
<tr>
<td>6.16</td>
<td>Miscellaneous Occasionally Useful Tidbits</td>
<td>137</td>
</tr>
<tr>
<td>6.16.1</td>
<td>Specifying directories</td>
<td>137</td>
</tr>
<tr>
<td>6.16.2</td>
<td>MAXIMA's Default Directory</td>
<td>137</td>
</tr>
<tr>
<td>6.16.3</td>
<td>MAXIMA's Search Path</td>
<td>138</td>
</tr>
<tr>
<td>6.16.4</td>
<td>Customizing MAXIMA</td>
<td>139</td>
</tr>
<tr>
<td>6.16.5</td>
<td>Restoring MAXIMA's Initial State</td>
<td>139</td>
</tr>
<tr>
<td>6.16.6</td>
<td>Placing Text in Graphs</td>
<td>139</td>
</tr>
<tr>
<td>6.16.7</td>
<td>Space Curves</td>
<td>140</td>
</tr>
<tr>
<td>6.17</td>
<td>Miscellaneous Commands</td>
<td>140</td>
</tr>
<tr>
<td>6.18</td>
<td>References</td>
<td>141</td>
</tr>
<tr>
<td>6.19</td>
<td>Exercises</td>
<td>142</td>
</tr>
</tbody>
</table>

## 9 Introduction to Programming 153

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.1</td>
<td>Components of a Programming Language</td>
<td>154</td>
</tr>
<tr>
<td>9.1.1</td>
<td>Variables, Variable Names, and Use of Memory</td>
<td>154</td>
</tr>
<tr>
<td>9.1.2</td>
<td>Essential Action Statements</td>
<td>156</td>
</tr>
<tr>
<td>9.1.3</td>
<td>Logical Expressions and Conditions</td>
<td>157</td>
</tr>
<tr>
<td>9.1.4</td>
<td>Syntactic Wrinkles</td>
<td>158</td>
</tr>
<tr>
<td>9.1.5</td>
<td>Essential Control Structures</td>
<td>158</td>
</tr>
<tr>
<td>9.2</td>
<td>Sample Short Algorithms</td>
<td>162</td>
</tr>
<tr>
<td>9.3</td>
<td>Two Larger Algorithms</td>
<td>164</td>
</tr>
<tr>
<td>9.3.1</td>
<td>Solving Laplace’s Equation</td>
<td>164</td>
</tr>
<tr>
<td>9.3.2</td>
<td>File Output/Input</td>
<td>166</td>
</tr>
<tr>
<td>9.9</td>
<td>Solving Laplace’s Equation with MATLAB</td>
<td>169</td>
</tr>
<tr>
<td>9.12</td>
<td>Creating and Storing Two-Dimensional Scalar Arrays</td>
<td>170</td>
</tr>
<tr>
<td>9.12.4</td>
<td>... with MATLAB</td>
<td>170</td>
</tr>
<tr>
<td>9.13</td>
<td>Creating and Storing Three-Dimensional Scalar Arrays</td>
<td>171</td>
</tr>
<tr>
<td>9.13.4</td>
<td>... with MATLAB</td>
<td>172</td>
</tr>
<tr>
<td>9.14</td>
<td>Creating and Storing Two-Dimensional Vector Arrays</td>
<td>172</td>
</tr>
<tr>
<td>9.14.4</td>
<td>... with MATLAB</td>
<td>174</td>
</tr>
<tr>
<td>9.15</td>
<td>Creating and Storing Three-Dimensional Vector Arrays</td>
<td>175</td>
</tr>
<tr>
<td>9.15.4</td>
<td>... with MATLAB</td>
<td>175</td>
</tr>
<tr>
<td>9.16</td>
<td>Reading Files</td>
<td>176</td>
</tr>
<tr>
<td>9.16.4</td>
<td>... with MATLAB</td>
<td>176</td>
</tr>
<tr>
<td>9.17</td>
<td>References</td>
<td>177</td>
</tr>
<tr>
<td>9.18</td>
<td>Exercises</td>
<td>177</td>
</tr>
<tr>
<td>9.F</td>
<td>Listing of \texttt{lugen_list.m} (\textsc{MATLAB})</td>
<td>183</td>
</tr>
<tr>
<td>9.R</td>
<td>Listing of \texttt{luread_list.m} (\textsc{MATLAB})</td>
<td>185</td>
</tr>
</tbody>
</table>
### 11 Solving Ordinary Differential Equations

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>11.1 Sample Problems</strong></td>
<td>187</td>
</tr>
<tr>
<td>11.1.1 Projectile in a Viscous Medium</td>
<td>188</td>
</tr>
<tr>
<td>11.1.2 Chain Radioactive Decay</td>
<td>191</td>
</tr>
<tr>
<td>11.1.3 Exponential and Logistic Growth</td>
<td>192</td>
</tr>
<tr>
<td>11.1.4 Forced, Driven, Damped Harmonic Oscillation</td>
<td>192</td>
</tr>
<tr>
<td>11.1.5 An LRC Resonant Circuit</td>
<td>194</td>
</tr>
<tr>
<td>11.1.6 Coupled Oscillators</td>
<td>194</td>
</tr>
<tr>
<td>11.1.7 Motion under Central Forces</td>
<td>195</td>
</tr>
<tr>
<td>11.1.8 Standing Waves in a String</td>
<td>197</td>
</tr>
<tr>
<td>11.1.9 The Schrödinger Equation in One Dimension</td>
<td>198</td>
</tr>
<tr>
<td><strong>11.2 Laplace Transforms</strong></td>
<td>199</td>
</tr>
<tr>
<td><strong>11.3 Solving ODEs Symbolically with MAXIMA</strong></td>
<td>200</td>
</tr>
<tr>
<td>11.3.1 Projectile in a Viscous Medium</td>
<td>202</td>
</tr>
<tr>
<td>11.3.2 Logistic Growth</td>
<td>204</td>
</tr>
<tr>
<td>11.3.3 Damped Harmonic Oscillator</td>
<td>206</td>
</tr>
<tr>
<td>11.3.4 Chain Radioactive Decay</td>
<td>206</td>
</tr>
<tr>
<td>11.3.5 Coupled Oscillators</td>
<td>208</td>
</tr>
<tr>
<td>11.3.6 Standing Waves in a String</td>
<td>209</td>
</tr>
<tr>
<td>11.3.7 Infinite Depth Quantum Well</td>
<td>210</td>
</tr>
<tr>
<td>11.3.8 Finding Solutions by Series Methods</td>
<td>211</td>
</tr>
<tr>
<td><strong>11.6 Algorithms for Solving ODEs Numerically</strong></td>
<td>213</td>
</tr>
<tr>
<td>11.6.1 Euler’s Method</td>
<td>214</td>
</tr>
<tr>
<td>11.6.2 Improved Euler Method</td>
<td>216</td>
</tr>
<tr>
<td>11.6.3 Runge-Kutta Methods</td>
<td>217</td>
</tr>
<tr>
<td>11.6.4 Assessing Accuracy</td>
<td>219</td>
</tr>
<tr>
<td>11.6.5 Adaptive Methods</td>
<td>221</td>
</tr>
<tr>
<td>11.6.6 Multistep Methods</td>
<td>222</td>
</tr>
<tr>
<td><strong>11.8 Solving ODEs Numerically with MATLAB</strong></td>
<td>222</td>
</tr>
<tr>
<td>11.8.1 Using Elementary Commands</td>
<td>223</td>
</tr>
<tr>
<td>11.8.2 Defining ODEs for MATLAB</td>
<td>226</td>
</tr>
<tr>
<td>11.8.3 The Procedures <code>ode23</code>, <code>ode45</code>, and <code>ode113</code></td>
<td>226</td>
</tr>
<tr>
<td>11.8.4 Chain Radioactive Decay</td>
<td>227</td>
</tr>
<tr>
<td>11.8.5 Damped Harmonic Oscillator</td>
<td>230</td>
</tr>
<tr>
<td>11.8.6 Planetary Orbits</td>
<td>231</td>
</tr>
<tr>
<td>11.8.7 Standing Waves in a String</td>
<td>234</td>
</tr>
<tr>
<td>11.8.8 The Quantum Harmonic Oscillator</td>
<td>237</td>
</tr>
<tr>
<td>11.8.9 Changing Tolerances</td>
<td>241</td>
</tr>
<tr>
<td>11.8.10 M-files in the Public Library</td>
<td>243</td>
</tr>
<tr>
<td><strong>11.11 Solving ODEs Numerically with MAXIMA</strong></td>
<td>244</td>
</tr>
<tr>
<td><strong>11.16 Exercises</strong></td>
<td>246</td>
</tr>
<tr>
<td>11.16.1 ... using Symbolic Methods</td>
<td>246</td>
</tr>
<tr>
<td>11.16.2 ... using Numerical Methods</td>
<td>249</td>
</tr>
</tbody>
</table>

### 12 Evaluating Integrals

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>13.1 Sample Problems</strong></td>
<td>255</td>
</tr>
<tr>
<td>13.1.1 One-Dimensional Trajectories</td>
<td>255</td>
</tr>
<tr>
<td>13.1.2 Center of Mass</td>
<td>256</td>
</tr>
<tr>
<td>13.1.3 Moment of Inertia; Radius of Gyration</td>
<td>257</td>
</tr>
<tr>
<td>13.1.4 The Large Amplitude Simple Pendulum</td>
<td>258</td>
</tr>
<tr>
<td>13.1.5 Statistical Data Analysis</td>
<td>259</td>
</tr>
<tr>
<td>13.1.6 The Cornu Spiral</td>
<td>260</td>
</tr>
<tr>
<td>13.1.7 Electric and Magnetic Fields and Potentials</td>
<td>260</td>
</tr>
<tr>
<td>13.1.8 Quantum Probabilities</td>
<td>262</td>
</tr>
<tr>
<td>13.1.9 Expansion in Orthogonal Functions</td>
<td>262</td>
</tr>
<tr>
<td><strong>13.2 Evaluating Integrals Symbolically with MAXIMA</strong></td>
<td>263</td>
</tr>
<tr>
<td>13.2.1 Relativistic Motion Under a Constant Force</td>
<td>264</td>
</tr>
<tr>
<td>13.2.2 Center of Mass</td>
<td>265</td>
</tr>
<tr>
<td>13.2.3 Moment of Inertia; Radius of Gyration</td>
<td>265</td>
</tr>
<tr>
<td>13.2.4 Electrostatic Potential of a Finite Line Charge</td>
<td>266</td>
</tr>
</tbody>
</table>
14 Finding Roots

14.1 Sample Problems ............................................. 305
14.1.1 Classical Turning Points ................................. 306
14.1.2 A Max-Min Problem: Equilibrium Points ............ 306
14.1.3 Natural Frequencies of Oscillating Systems ....... 307
14.1.4 Range of Projectile in Viscous Medium ............ 308
14.1.5 Energy Levels in a Quantum Well .................... 311

14.2 Symbolic Approaches ........................................... 312

14.3 Finding Roots Symbolically with MAXIMA .................. 313
14.3.1 Classical Turning Points ................................. 313
14.3.2 Equilibrium Points ................................... 314
14.3.3 Natural Frequencies .............................. 315
14.3.4 Range of Projectile in Viscous Medium ............ 315

14.4 Algorithms for Finding Roots Numerically ............... 318
14.4.1 The Method of Bisection .............................. 318
14.4.2 Newton’s Method .................................... 319
14.4.3 Other Methods ........................................ 319
14.4.4 Assessing Accuracy .................................... 320

14.8 Finding Roots Numerically with MATLAB .................. 321
14.8.1 Classical Turning Points ................................. 322
14.8.2 Natural Frequencies .............................. 323
14.8.3 Range of Projectile ................................ 324
14.8.4 Energy Levels in Quantum Well ..................... 327

14.11 Finding Roots Numerically with MAXIMA ............... 329
14.11.1 Classical Turning Points ................................. 331
14.11.2 Range of Projectile ................................ 332
14.11.3 Energy Levels in Quantum Well ..................... 333

14.16 Solving Simultaneous Equations ......................... 335
14.16.1 Systems of Linear Equations ......................... 336
14.16.2 Systems of Nonlinear Equations .................... 339

14.17 Exercises ...................................................... 341
14.17.1 ... using Symbolic Methods ......................... 341
14.17.2 ... using Numerical Methods ....................... 343
15 Solving Partial Differential Equations

15.1 Sample Problems
15.1.1 Motion of a String: The Wave Equation
15.1.2 Heat Flow: The Diffusion Equation
15.1.3 Steady State Heat Flow: The Laplace Equation
15.1.4 Other Situations
15.1.5 Classification of Second-Order PDEs
15.1.6 Equations of Fluid Dynamics: Navier-Stokes’ Equations
15.1.6.1 Conservation of Mass: The Equation of Continuity
15.1.6.2 Newton’s Second Law
15.1.6.3 A Special Case: Non-viscous Flow
15.1.6.4 A Second Special Case: Sound Waves
15.1.7 A General 1D Equation
15.1.8 A General 2D Equation

15.2 Finite Difference Methods (FDMs) in One Dimension
15.4 Using MATLAB to Solve 1D PDEs via an FDM
15.4.1 A General Coding
15.4.2 An Example: Simple Harmonic Motion

15.9 Finite Element Methods (FEMs) in One Dimension
15.9.1 Discretizing the Domain: Preprocessing
15.9.2 Selecting Interpolation or Shape Functions
15.9.3 Formulating the Equations for a Single Element
15.9.4 Assembling the System of Equations
15.9.5 Incorporating the Boundary Conditions

15.11 Using MATLAB to Solve 1D PDEs via an FEM
15.11.1 A General Coding
15.11.2 An Example: Simple Harmonic Motion

15.16 Finite Difference Methods (FDMs) in Two Dimensions
15.16.1 The Wave Equation
15.16.2 The Diffusion Equation
15.16.3 The Laplace Equation

15.18 Using MATLAB to Solve 2D PDEs via an FDM
15.18.1 The Wave Equation
15.18.2 The Diffusion Equation
15.18.3 The Laplace Equation

15.23 Finite Element Methods (FEMs) in Two Dimensions
15.23.1 Discretizing the Domain: Pre-processing
15.23.2 Selecting the Interpolation or Shape Functions
15.23.3 Formulating the Equations for a Single Element
15.23.4 Assembling the System of Equations
15.23.5 Incorporating the Boundary Conditions

15.25 Using MATLAB to Solve 2D PDEs via an FEM
15.25.1 A General Coding
15.25.2 An Example: Isotropic Heat Flow

15.30 Exercises

15.A Program(s) for FDM Approach to 1D Problem
15.A.2 Listing of fdmlap.m (MATLAB)

15.B Evaluating Integrals for 1D FEM Problem
15.B.1 . . . using MAXIMA

15.C Program(s) for FEM Approach to 1D Problem
15.C.2 Listing of femlap.m (MATLAB)

15.D MATLAB Programs for FDM Approach to 2D Problems
15.D.1 Listing of fdwave.m
15.D.2 Listing of fdmsdiff.m
15.D.3 Listing of fdmslap.m
15.D.4 Listing of fdmslap2d.m

15.E A Useful Integral
TABLE OF CONTENTS

xvi

A Introduction to \LaTeX

A.1 Creating a Simple Document .................................................. 458
  A.1.1 Structuring a \LaTeX\ Source File ..................................... 458
  A.1.2 Creating a PostScript File ............................................ 460
  A.1.3 Creating a PDF File .................................................... 461
  A.1.4 Displaying the Document on the Screen ............................. 462
  A.1.5 Printing the Document ................................................ 462
A.2 Specification of Global Style: The Preamble ............................... 463
A.3 In-Text Specification of Local Style ....................................... 465
  A.3.1 Type Style .................................................................. 465
  A.3.2 Type Size .................................................................. 466
  A.3.3 White Space .............................................................. 467
  A.3.4 Environments ................................................................ 468
  A.3.5 \LaTeX\ Packages ........................................................ 469
  A.3.6 Miscellaneous Other Capabilities .................................... 469
A.4 Including Equations ................................................................ 470
A.5 Including Lists ......................................................................... 471
A.6 Including Tables ....................................................................... 472
A.7 Including Illustrations ............................................................. 474
  A.7.1 Using Cut and Paste ...................................................... 474
  A.7.2 Using the graphicx Package ............................................ 475
  A.7.3 Using the tikz Package .................................................. 477
  A.7.4 Using the picture Environment and pict2e Package ............ 480
A.8 Including TOC, LOF, and LOT .................................................... 480
A.9 Including an Index ................................................................. 481
A.10 Including Hyperlinks ............................................................... 483
A.11 Converting .eps and .ps Files to .pdf ........................................ 484
A.12 Using Conditional Expressions in \LaTeX ................................. 486
A.13 Error Messages Generated by \LaTeX ......................................... 487
A.14 The Page Previewer for .dvi Files ........................................... 488
A.15 The Spell Checker in UNIX .................................................. 489
A.16 A Sample Document ............................................................... 490
A.17 Miscellaneous Other Features ................................................ 491
A.18 References ............................................................................. 498
A.19 Exercises ................................................................................. 499
A.A Listings ....................................................................................... 502
  A.A.1 . . . for Windows ............................................................. 502
    A.A.1.1 Batch File ceps2pdf.bat ........................................... 502
    A.A.1.2 Batch File cps2pdf.bat ............................................ 502
    A.A.1.3 Batch File rdfileeps.bat ......................................... 502
    A.A.1.4 Batch File rdfileps.bat ........................................... 502
  A.A.2 . . . for UNIX ................................................................. 503
    A.A.2.1 Shell Script ceps2pdf .............................................. 503
    A.A.2.2 Shell Script cps2pdf ............................................... 503
    A.A.2.3 Shell Script rdfileeps ............................................. 503
    A.A.2.4 Shell Script rdfileps ............................................... 503
  A.A.3 . . . for Windows and UNIX ................................................ 504
    A.A.3.1 Python Script ExtractFileName.py ............................ 504

Z Contacting Software Vendors ......................................................... 505

Index .............................................................................................. 509
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., \TeX\Live, \LaTeX, Mi\TeX, Oz\TeX, ...) capable of formatting elaborate equations, incorporating symbolic references within documents, generating tables of contents and indices, ....

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

\footnote{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. AN ORIENTATION TO COMPUTERS

1.1.1 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

\[\text{The special words LOOP and END LOOP bracket a group or block of instructions that are as a block to be executed repeatedly. We shall here ignore concerns about stopping the loop.}\]

\[\text{The special words IF, THEN, and END_IF convey a conditional execution of one or more statements. The statement(s) between the THEN and the END_IF will be executed only if the condition following the IF is true when the entire construction is encountered.}\]
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,
\[ 1 \times 128 + 0 \times 64 + 0 \times 32 + 1 \times 16 + 1 \times 8 + 1 \times 4 + 0 \times 2 + 1 \times 1 = 157 \]

The number eight is, of course, arbitrary but conventional. Because computers work internally in the binary (base-2) number system, powers of two—8 = 2^3—are especially convenient.

Actually, the codes will be sent as a stream of bits, each of which is an electrical voltage level that will be either “high” or “low”, often said to be “on” or “off”.

5 The number eight is, of course, arbitrary but conventional. Because computers work internally in the binary (base-2) number system, powers of two—8 = 2^3—are especially convenient.

6 Actually, the codes will be sent as a stream of bits, each of which is an electrical voltage level that will be either “high” or “low”, 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 nibbles. 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.
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 \(2^{16} = 65536\) 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 \(2^{32} = 4294967296\) 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>DEC</th>
<th>CHAR</th>
<th>DEC</th>
<th>CHAR</th>
<th>DEC</th>
<th>CHAR</th>
<th>DEC</th>
<th>CHAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>NUL</td>
<td>16</td>
<td>DLE</td>
<td>32</td>
<td>SP</td>
<td>48</td>
<td>@</td>
</tr>
<tr>
<td>1</td>
<td>SOH</td>
<td>17</td>
<td>DC1</td>
<td>33</td>
<td>!</td>
<td>49</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>STX</td>
<td>18</td>
<td>DC2</td>
<td>34</td>
<td>&quot;</td>
<td>50</td>
<td>B</td>
</tr>
<tr>
<td>3</td>
<td>EOX</td>
<td>19</td>
<td>DC3</td>
<td>35</td>
<td>#</td>
<td>51</td>
<td>C</td>
</tr>
<tr>
<td>4</td>
<td>EOT</td>
<td>20</td>
<td>DC4</td>
<td>36</td>
<td>$</td>
<td>52</td>
<td>D</td>
</tr>
<tr>
<td>5</td>
<td>ENQ</td>
<td>21</td>
<td>NAK</td>
<td>37</td>
<td>%</td>
<td>53</td>
<td>E</td>
</tr>
<tr>
<td>6</td>
<td>ACK</td>
<td>22</td>
<td>SYN</td>
<td>38</td>
<td>'</td>
<td>54</td>
<td>F</td>
</tr>
<tr>
<td>7</td>
<td>BEL</td>
<td>23</td>
<td>ETB</td>
<td>39</td>
<td>'</td>
<td>55</td>
<td>G</td>
</tr>
<tr>
<td>8</td>
<td>BS</td>
<td>24</td>
<td>CAN</td>
<td>40</td>
<td>(</td>
<td>56</td>
<td>H</td>
</tr>
<tr>
<td>9</td>
<td>HT</td>
<td>25</td>
<td>EM</td>
<td>41</td>
<td>)</td>
<td>57</td>
<td>I</td>
</tr>
<tr>
<td>10</td>
<td>LF</td>
<td>26</td>
<td>SUB</td>
<td>42</td>
<td>*</td>
<td>58</td>
<td>J</td>
</tr>
<tr>
<td>11</td>
<td>VT</td>
<td>27</td>
<td>ESC</td>
<td>43</td>
<td>+</td>
<td>59</td>
<td>K</td>
</tr>
<tr>
<td>12</td>
<td>FF</td>
<td>28</td>
<td>FS</td>
<td>44</td>
<td>,</td>
<td>60</td>
<td>L</td>
</tr>
<tr>
<td>13</td>
<td>CR</td>
<td>29</td>
<td>GS</td>
<td>45</td>
<td>-</td>
<td>61</td>
<td>M</td>
</tr>
<tr>
<td>14</td>
<td>SO</td>
<td>30</td>
<td>RS</td>
<td>46</td>
<td>.</td>
<td>62</td>
<td>N</td>
</tr>
<tr>
<td>15</td>
<td>SI</td>
<td>31</td>
<td>US</td>
<td>47</td>
<td>/</td>
<td>63</td>
<td>O</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 `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.\(^{10}\) Some files—called ASCII text files—contain nothing but printing ASCII characters (and perhaps such simple control characters as \(<\text{CR}>\) and \(<\text{HT}>\)) 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, \textit{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.\(^{11}\) Locating a specific file in the entire structure then requires not only giving the file name but also describing its \textit{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 `/`; 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

```
/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}.\(^{12}\) 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.

\(^{10}\)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.

\(^{11}\)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.

\(^{12}\)In Windows, for example, the backslash character \texttt{\} 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.\textsuperscript{13} The Local Guide explains how to translate these symbols if your mouse has fewer than three buttons.

\textsuperscript{13}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.
• 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 WinEdt 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

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

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

• 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. \texttt{x1} or \texttt{xf}, in the statement to be executed but render these variables as subscripted, e.g., \texttt{x_1} or \texttt{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...
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

\[
10!! = 10 \times 8 \times 6 \times 4 \times 2 = 2^5 \times 5 \times 4 \times 3 \times 2 \times 1 = 2^5 \cdot 5!
\]

\[\implies (2n)!! = 2^n n! \tag{1.3}\]

If \( n \) is odd, say 9, recasting its double factorial takes mildly more work, but is illustrated in the chain

\[
9!! = 9 \times 7 \times 5 \times 3 \times 1 = \frac{9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{8 \times 6 \times 4 \times 2} = \frac{9!}{8!!} = \frac{9!}{2^4 4!}
\]

\[\implies (2n + 1)!! = \frac{(2n + 1)!}{2^n n!} \tag{1.4}\]

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

\[
\left[\Gamma\left(\frac{1}{2}\right)\right]^2 = \left(\int_{0}^{\infty} e^{-x^2} \, dx\right)^2 = \left(\int_{0}^{\infty} e^{-x^2} \, dx\right) \left(\int_{0}^{\infty} e^{-y^2} \, dy\right)
\]

\[
= \int_{0}^{\infty} \int_{0}^{\infty} e^{-(x^2+y^2)} \, dx \, dy = \int_{0}^{\pi} e^{-r^2} r \, dr \int_{0}^{2\pi} d\phi = \pi \tag{1.8}
\]
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
\[
\Gamma(\nu) = -\int_{0}^{\infty} t^{\nu-1} e^{-t} \big|_0^\infty + (\nu - 1) \int_{0}^{\infty} t^{\nu-2} e^{-t} dt = (\nu - 1) \Gamma(\nu - 1) \] (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
\[
\mathcal{L}(f(t)) = \tilde{f}(s) = \int_{0}^{\infty} e^{-st} f(t) \, dt \] (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,
\[
\mathcal{L} \left( a f(t) + b g(t) \right) = \int_{0}^{\infty} e^{-st} \left( a f(t) + b g(t) \right) \, dt = a \int_{0}^{\infty} e^{-st} f(t) \, dt + b \int_{0}^{\infty} e^{-st} g(t) \, dt = a \mathcal{L} \left( f(t) \right) + b \mathcal{L} \left( g(t) \right) \] (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 \( \nu > 1 \).
CHAPTER 1. PRELIMINARIES

Table 1.2: A short table of Laplace transforms.

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

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

- 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 3

Introduction to MATLAB

*Note:* All program (*.m) and data (*.dat) files referred to in this chapter are available in the directory `$HEAD/matlab`, where (as defined in the Local Guide) `$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 MATLAB's default search path. (See Section 3.18.3.) If so, the files can be found by MATLAB 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.

MATLAB®—an acronym for Matrix Laboratory—is an interactive software package 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), two-dimensional matrices (m \times n arrays), or higher dimensional arrays, and to produce graphical visualizations.¹ A typical interaction with the program will involve (1) creating one or more arrays in MATLAB’s workspace (either by computing them directly in MATLAB or by importing them into MATLAB 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 on-line help messages accessible from within MATLAB itself (see Section 3.8) and in assorted MATLAB documentation (see Section 3.19). We shall refer to this documentation collectively as the MATLAB manuals.

Two user interfaces to MATLAB are provided, a command-line interface (CLI) and a graphical user interface (GUI), called the MATLAB desktop, which facilitates working with files containing MATLAB programs. Statements to MATLAB are entered in a UNIX Shell window or a MATLAB Command window within the MATLAB desktop. Some of those statements, however, create additional windows within which a variety of actions can be selected with the mouse. With either interface, statements to MATLAB are structured and typed in the same way. Thus, in this chapter, we limit ourselves for the most part to describing the CLI.

3.1 Beginning a MATLAB Session

Detailed instructions for initiating a session with MATLAB will be found in the Local Guide. Usually, MATLAB will be started either (1) by typing the command `matlab` (for the GUI) or the command `matlab -nodesktop` (for the CLI) at the prompt from the operating system,² (2) by double-clicking the left mouse button on an appropriate icon on the desktop, or (3) by selecting MATLAB from a

¹MATLAB is a registered trademark of and also a commercially available program produced and marketed by The MathWorks, Inc. (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 The MathWorks, Inc. The terms of that license are explained in the Local Guide.

²In these commands to the operating system, case may be important.
Presently, the MATLAB prompt >> will appear, either in the Shell window in which the command was typed (CLI) or in a newly created MATLAB desktop, the most prominent component of which is a MATLAB Command Window (GUI). We must from the outset be aware that

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

2. Typing the command quit at the MATLAB prompt or selecting ‘Exit MATLAB’ from the File menu in the MATLAB Command window, or typing ⟨CONTROL-q⟩ ends the session.

3. (Control-C) instructs MATLAB to abort its present activity more or less immediately. Presently, the MATLAB prompt returns and another statement can be submitted for execution.

4. Every command line in MATLAB 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 MATLAB 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 percent sign ‘%’ can be used to introduce a comment. The percent sign and all following characters in the line will be ignored.

7. An initial exclamation point ‘!’ allows temporary escape from MATLAB to the operating system. Thus, for example, the statement ‘!lp file.ps’ (UNIX) or ‘!copy file.ps lpt1:’ (Windows) issued to MATLAB will be executed not by MATLAB but rather by the underlying operating system. Control is then returned to MATLAB. If you are working in the GUI, this feature is, of course, unnecessary—since we can easily use resources of the operating system’s desktop to execute other programs without exiting from MATLAB.

8. Within MATLAB, we can retrieve previous statements with the up-arrow key and edit individual statements by using the left and right arrow keys to move the cursor in the line, the backspace key to delete the character to the left of the cursor, and other keys to insert characters at the position of the cursor. Further, pressing RETURN will execute the entire command line, regardless of where the cursor happens to be positioned within the line.

For examples of MATLAB’s capabilities, select ‘Examples and Demos’ (or, in more recent versions, ‘Demos’) from the Help menu in the GUI or type demo at the MATLAB prompt. Then surf through the multitude of menus provided. More recent versions of MATLAB require Adobe Flash® Player® and an internet connection if you elect to display any videos.

MATLAB affords considerable customizability to the appearance of the desktop in the GUI, and you may wish to explore that capacity. For the sake of easy description, however, the text here assumes the default appearance of the desktop. That appearance will probably be what is presented to you when you first launch the GUI, but it can be established in any case by selecting ‘Default’ from the item ‘Desktop Layout’ from the Desktop menu at the top of the GUI.

3Behind the scenes at the start of a MATLAB session, the start-up script checks to see that the license manager is running and 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.

4In more recent versions of MATLAB, the label in that window includes an identification of the version in use.
3.2 Basic Entities in MATLAB

3.2.1 Data Types

While individual items of data can be aggregated into quite complicated structures, only two basic data types—the term used by MATLAB is classes—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 $10^{308}$. MATLAB refers to this data class as double. For character strings (class char), on the other hand, MATLAB 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. MATLAB 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 MATLAB commands num2str, int2str, and str2num, can be used to convert a numeric value to a string, an integer value to a string, and a string to a numeric value, respectively. In addition, the commands fix and round can be used to convert a floating point value into an integer value, where fix truncates the value at the decimal point while 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’, and the conversion of the floating point value 3.67 with the expression int2str(3.67), which will yield the string ‘4’, information will be lost in the rounding and cannot be retrieved by invoking the reverse command str2num.

While MATLAB 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`.

3.2.2 Variable Names

The simplest variables in MATLAB 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 up to and including the value returned by the command `namelengthmax`. Upper- and lower-case letters, numbers, and the underscore character are legal. 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.

3.2.3 Assignment of Values to Variables

In MATLAB, 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

5 A string value (e.g., ‘cook’) that does not represent a valid number cannot be converted to a number.
6 In the system used for the development of this book, that value by default is 63.
variable, i.e., stores in an eight-byte memory location named \texttt{alpha} the (double precision) floating point binary representation of the value 3.79. Similarly, the statement \texttt{name = 'David'} creates a variable of class \texttt{char} and assigns the value \texttt{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 quotation marks (double quotation marks will generate an error message) and (2) that the values returned by a MATLAB 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 MATLAB 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

\begin{verbatim}
x = alpha + beta
\end{verbatim}

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

### 3.2.4 Commands

In MATLAB 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 MATLAB's \textit{commands}, the behavior of which is almost always influenced by the value or values of one or more arguments and/or \textit{properties}. Statements invoking a particular command will usually be expressed in the form

\begin{verbatim}
command( argument, argument, 'PropertyName', PropertyValue, ...)
\end{verbatim}

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

\begin{verbatim}
q = command( argument, argument, 'PropertyName', PropertyValue, ...)
\end{verbatim}

in which the value returned by the command is assigned to the user-specified variable \texttt{q}—though MATLAB will assume the variable \texttt{ans} if no variable is specified. The order of the "free" arguments is mandatory, since the position of each argument identifies its role.\footnote{In those situations where no value is to be specified for a free argument but a value must be specified for one that follows it in the prescribed order, the symbol [] can be used as a placeholder for the argument to be unspecified.} 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 \texttt{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 MATLAB in the form \texttt{hold on}. Full information about this feature, which MATLAB refers to as command/function duality can be found in the MATLAB manuals.

3.3 A Sampling of MATLAB Capabilities

In this section, we present several examples illustrating various capabilities of MATLAB and introducing some of the most frequently used commands. The “conversation” in this section should start in a fresh invocation of MATLAB.

3.3.1 Creating and Examining Arrays

Arrays can be created within MATLAB or—see Section 3.7—read in from files. Individual elements of arrays can be accessed by specifying the row and column, as in \texttt{array\_name(row,column)}, where the \textit{first} index identifies the row, the \textit{second} index identifies the column, the indices are enclosed in parentheses, and they begin at one (repeat: \textit{indices begin at 1}).\footnote{We here limit our discussion to arrays of one and two dimensions. See the MATLAB manuals for ways to create arrays of three or more dimensions.} Alternatively, individual elements can be accessed by specifying the element number, as in \texttt{array\_name(element number)}. For this form, the elements in the array are numbered sequentially \textit{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\footnote{MATLAB statements are shown on the left; comments describing the statements are shown on the right. Further, to save space, we will routinely compact MATLAB’s output by omitting blank lines and extra spaces.}

\begin{verbatim}
>> name = 'David'
name = David

>> val = 3.1415926535
val = 3.1416

>> cmp = 3.0+5.0*i
cmp = 3.0000+5.0000i
\end{verbatim}

Assign string value to \texttt{name} and print it to the screen. The actual variable will be a $1 \times 5$ array of class \texttt{char}.

Assign numeric value to \texttt{val} and print it to the screen. The actual variable will be a $1 \times 1$ array of class \texttt{double}. All digits are stored internally. By default, for \textit{display}, MATLAB adopts its \texttt{short} format, rounding (decimal) numbers to four digits after the decimal point.

Assign a \texttt{complex} value to the variable \texttt{cmp}, using the \texttt{special symbol} \texttt{i} for $\sqrt{-1}$. MATLAB also recognizes the special symbol \texttt{j} for $\sqrt{-1}$.

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

\begin{verbatim}
>> b = [ 1 0 3 7 10 ]
b = 1 0 3 7 10
\end{verbatim}

Create a (row) vector \texttt{b}, separating elements in a row with spaces (commas would also work), and print it to the screen.
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. This time, we separate elements in a row with commas (or spaces); rows are separated by semicolons.

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.\(^{10}\)

Other functions that can be used to create arrays include `eye(n)` (to make an $n \times n$ unit matrix), `magic(n)` (to make an $n \times n$ magic square), and `pascal(n)` (to make an $n \times n$ matrix whose columns and rows are the Pascal numbers).

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

```
>> who
Your variables are:
   a  b  cmp  data  name  val  z
```

```
>> whos
Name      Size     Bytes    Class     Attributes
a 3x3       72        double
b 1x5       40        double
   Name  Size    Bytes   Class   Attributes
cmp  1x1       16        double complex
data 100x100   80000     double
   Name  Size    Bytes   Class
name  1x5      10        char
val   1x1      8         double
z 3x3       72        double
```

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')`]. Any one of the statements `clear`, `clear all`, or `clear('all')` would remove all known variables.

\(^{10}\)Typically, generators of what are technically called pseudorandom numbers calculate each new random number from the one just generated. When MATLAB is started, the initial "state" of the random number generator is fixed and will be the same with each new start. Then, each invocation of `rand` will change that state, and the next number to be generated will be determined from that new state. When debugging programs involving random numbers, we may well find it convenient now and then to reset the initial state with the statement `rand('state',0)`, thereby forcing MATLAB to generate again the same sequence of random numbers. Once a program has been debugged, however, we really don’t want each run to use the same sequence of random numbers. We should therefore begin any production run with a statement like `rand('state',100*sum(clock))`, which will set the state to a random initial value determined by the detailed current reading of the system clock and minimize the likelihood that two separate runs will generate the same sequence of random numbers.
Once an array has been created, we can access individual elements in the array with statements like

\[
\text{\texttt{>> a(1,2)}} \quad \text{Print the value in the first row, second column of } a. \text{ MATLAB uses the variable } \texttt{ans} \text{ if no “output” variable is specified.}
\]
\[
\text{\texttt{>> a(4)}} \quad \text{Print the fourth element of } a. \text{ Note that this element is the same as the element printed by the previous statement, but it is being addressed by element number rather than by its column and row indices.}
\]

\[
\text{\texttt{>> a(4) = 10}} \quad \text{Assign the value 10 to the fourth element (first row, second column) of } a. \text{ The edited array is printed automatically.}
\]
\[
\text{\texttt{>> a(2,:)}} \quad \text{Print the second row of } a. \text{ The colon by itself functions as a wild card and stands for all values of the corresponding index.}
\]
\[
\text{\texttt{>> c = a(1:2,1:2)}} \quad \text{Extract a } 2 \times 2 \text{ matrix made from the four elements in the upper left hand corner of } a \text{ and print it. The colon between two values tells MATLAB to use a range of values running in integer steps from the first to the last, inclusive.}
\]

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

\[
\text{\texttt{>> format long}} \quad \text{Specify the long format and redisplay the value by asserting its name. This time 15 digits will be displayed. Here we use \texttt{format} as a command. We expose MATLAB’s command/function duality by noting that this statement might also have been written \texttt{format(’long’)}.}
\]
\[
\text{\texttt{>> format(’short’)}} \quad \text{Return to default format, this time exploiting command/function duality to use \texttt{format} as a function.}
\]

Other formats include

- \texttt{short e}, which forces presentation of numbers in short format but uses the scientific format (3.1415926535 becomes 3.1416e+00),
- \texttt{long e}, which forces presentation of numbers in long format but uses scientific notation,
- \texttt{short g}, which gives MATLAB the flexibility to use either \texttt{short} format or \texttt{short e} format, whichever requires fewer characters, and
- \texttt{long g}, which gives MATLAB the flexibility to use either \texttt{long} format or \texttt{long e} format, whichever requires fewer characters.

Regardless of the \textit{display}, internally MATLAB uses full double precision for all calculations and storage.
3.3.2 Simple Mathematical Manipulations

The ordinary mathematical operations of combining variables and constants are achieved with the usual operators: + for addition, - for subtraction, * for multiplication, / for division, and ^ for exponentiation. Within MATLAB, 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 ab^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, MATLAB 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 MATLAB’s operators .*, .^, and ./.

Further, MATLAB 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. The commands help elfun (elementary functions) and help specfun (special functions) will return lists of all functions in the corresponding category. Functions may have one or more arguments (see the MATLAB manuals). Those that have a single argument are invoked with a statement of the form

\[ q = \text{function}_\text{name}(x) \]

after the execution of which the variable q will contain the value of the corresponding function at the argument contained in the variable x.\(^{11}\) 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

\[ q = \text{atan2}(y, x) \]

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 3.3.1, the following record of a MATLAB session demonstrates a few mathematical manipulations:

```
>> eq1 = 8.4*sin(30.0*pi/180.0)*sqrt(3.0)
eq 1 = 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
```

\(^{11}\)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.
3.3. A SAMPLING OF MATLAB CAPABILITIES

Multiply \( c \) and \( d \) \textit{element by element} (using the operator \( .* \)).

Divide \( c \) by \( d \) \textit{element by element}. Note that \textit{floating} division has been done even though the elements in the matrix had no explicit decimal points.

3.3.3 Matrix Manipulations

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

\[
e = dc \quad \Rightarrow \quad e_{ij} = \sum_k d_{ik} \cdot c_{kj}
\]

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

Find the \textit{matrix product} \( dc \).

Find the \textit{matrix product} \( cd \), and note that \( dc \neq cd \). Matrix multiplication is, in general, not commutative.

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

Evaluate the determinant of \( c \).

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

Evaluate the transpose of \( c \).

Evaluate the inverse of \( c \).

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

Create two vectors, suppressing display.

Evaluate dot product, seeing it as the product of a row vector into a column vector.

Evaluate dot product, invoking built-in function \textit{dot}.

Since the two vectors \( a \) and \( b \) are created as \textit{row} vectors, the second factor in the dot product must be transposed to convert it into a column vector \textit{before} the matrix product is evaluated. As one might expect, MATLAB also has the analogous function \textit{cross} for cross products, though that operation is meaningful only for vectors with three components.
3.3.4 Solving Linear Equations

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

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

or

\[
\begin{pmatrix}
2 & 5 & 3 \\
-1 & 3 & -4 \\
1 & -1 & 0
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
x_3
\end{pmatrix}
=
\begin{pmatrix}
3 \\
-4 \\
1
\end{pmatrix}
\]

(3.2)

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 5 3; -1 3 -4; 1 -1 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 the equation \( a\mathbf{x} = \mathbf{b} \), where \( a \) is a square matrix and \( \mathbf{x} \) and \( \mathbf{b} \) are vectors, implies that \( \mathbf{x} = a^{-1}\mathbf{b} \), we might—though (from the perspective of numerical accuracy) almost always unwisely—seek a solution with the statements

```matlab
>> ainv = inv(a);
>> x = ainv * b
x =
0.8529
-0.1471
 0.6765
```

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

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

```matlab
>> x = a\b
x =
0.8529
-0.1471
 0.6765
```

yields the same result.

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

\[
\begin{pmatrix}
x_1 & x_2 & x_3
\end{pmatrix}
\begin{pmatrix}
2 & -1 & 1 \\
5 & 3 & -1 \\
3 & -4 & 0
\end{pmatrix}
=
\begin{pmatrix}
3 \\
-4 \\
 1
\end{pmatrix}
\]

(3.3)

i.e., in the form \( xc=\mathbf{b} \), where \( x \) and \( \mathbf{b} \) are now row vectors (and \( c—\text{notice}—\)is the transpose of \( a \)). In this case, the solution would be found by evaluating \( \mathbf{x}=bc^{-1} \), i.e. by multiplying \( \mathbf{b} \) from the right by the inverse of \( c \) (or dividing \( \mathbf{b} \) from the right by \( c \)). The MATLAB statements
yield the result we have already obtained, though the vectors are now row vectors rather than column vectors.\(^{12}\)

### 3.3.5 A First Graph

MATLAB’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, MATLAB uses the colon operator ‘:’ for this purpose. For example, we might create simple vectors with the statements

```matlab
>> y = [ 0 : 1 : 5 ]
y = 0 1 2 3 4 5

>> yy = [ -3 : 2 : 6 ]
yy = -3 -1 1 3 5
```

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

```matlab
>> x = [ 0.0 : 0.1 : 10.0 ];
```

or we might create the same vector using the MATLAB function `linspace` in a statement like

```matlab
>> x = linspace( 0.0, 10.0, 101 );
```

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, including the two bounding points. Thus, the number of segments into which the interval is divided is one 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 MATLAB statements

```matlab
>> dx = 6.0 / 100.0;
```

```matlab
>> x = [ -3.0 : dx : 3.0 ];
```

\[^{12}\text{Rather than entering the matrix \(c\) and the (row) vector \(b\) explicitly, we could alternatively have invoked the statements \(c = transpose(a)\) and \(b = transpose(b)\) to generate them from values already created.}\]
CHAPTER 3. INTRODUCTION TO MATLAB

Figure 3.1: The hyperbolic cosine function.

>> y = cosh( x );
Assign to each element of a vector y the hyperbolic cosine of the corresponding element of x.

>> plot( x, y )
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 command plot is more fully discussed in Section 3.11.

The graph created by these statements appears in a new window called a Figure window, a copy of which is shown in Fig. 3.1.

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

- use the command axis with a four-component vector as argument to specify the desired ranges \((x_{\text{min}}, x_{\text{max}})\) on the x axis and \((y_{\text{min}}, y_{\text{max}})\) on the y axis.
- use the command title to place a title on the graph.
- use the commands xlabel and ylabel to put labels on the x and y axes.
- use the command grid to replace short tics along the axes with full grid lines.

Further, we might stipulate values for any of a wide variety of MATLAB properties to customize the line weight and line style, to change character sizes, ... Alternatively, we can adjust some of these properties by selecting items from various menus in the Figure window in which the graph is presented on the screen.

Exploiting the additional commands and properties described in the previous paragraph, we might therefore produce the more readable and useful graph shown in Fig. 3.2 with the statements\(^{13}\)

\(^{13}\)Note the order of these statements. The original graph must be produced before any of the other embellishments can be added.
3.3. A SAMPLING OF MATLAB CAPABILITIES

Figure 3.2: The hyperbolic cosine function.

>> plot( x, y, 'Color', 'black', 'LineWidth', 4 )
>> 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 )

In these statements, each of the pairs 'Color', 'black', 'LineWidth', 4, 'FontSize', 20, and 'FontSize', 16 assigns the value given by its second member to the property identified by its first member. Specifically, they override the default color (blue) for the graph with black, the default linewidth (0.5 pt) with 4 pt, and the default character size (10 pt) with 20 pt or 16 pt.

The stipulation of color for a graph can be complicated. Here, we start a discussion to which we will return in Section 3.17.5. Table 3.1 shows several common colors that can be requested by a simple name and the three-component vector that stipulates the intensities of the RGB (red-green-blue) components of each. We could, for example, stipulate a black graph in the above statement with the variant

>> plot( x, y, 'Color', [0,0,0], 'LineWidth', 4 )

in which we specify a three-component vector giving the RGB components of the color and, clearly, introduce a means by which colors other than the standard eight can be produced. Alternatively, we can use the abbreviation

>> plot( x, y, 'Color', 'k', 'LineWidth', 4 )

or, even more simply, we can invoke a further shorthand and omit the property name altogether with the statement

>> plot( x, y, 'k', 'LineWidth', 4 )

though, in this case, the argument specifying the color must come before the stipulation of other properties.
Table 3.1: Color designations in MATLAB.

<table>
<thead>
<tr>
<th>Color Name</th>
<th>RGB Vector</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>black</td>
<td>[0,0,0]</td>
<td>k</td>
</tr>
<tr>
<td>blue</td>
<td>[0,0,1]</td>
<td>b</td>
</tr>
<tr>
<td>green</td>
<td>[0,1,0]</td>
<td>g</td>
</tr>
<tr>
<td>cyan</td>
<td>[0,1,1]</td>
<td>c</td>
</tr>
<tr>
<td>red</td>
<td>[1,0,0]</td>
<td>r</td>
</tr>
<tr>
<td>magenta</td>
<td>[1,0,1]</td>
<td>m</td>
</tr>
<tr>
<td>yellow</td>
<td>[1,1,0]</td>
<td>y</td>
</tr>
<tr>
<td>white</td>
<td>[1,1,1]</td>
<td>w</td>
</tr>
</tbody>
</table>

Note, incidentally, the power of the statement $y = \cosh(x)$. The argument of the function \texttt{cosh} is a vector and contains several values. In response to this statement, MATLAB 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. MATLAB automatically understands that intention with this simpler statement. Beyond simpler coding, MATLAB also achieves faster execution than would be achieved in more traditional coding.

3.4 Properties, Objects, and Handles

In the example in Section 3.3.5, we encountered the notion of properties. From the perspective of object-oriented programming, the creation of a graph in MATLAB entails the creation of several objects and the placement of these objects at appropriate points on the screen. The command \texttt{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—\texttt{plot} creates an axes object, within which it creates a line object—the graph itself. The commands \texttt{title}, \texttt{xlabel}, and \texttt{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 properties \texttt{Color}, \texttt{LineWidth}, and \texttt{FontSize} are among those which the user can change in this way.\footnote{In a departure from MATLAB's normal case-sensitive feature, the property names in the contexts of these commands are not case sensitive. Conventionally the first letter of each new word in the name is set in upper case to facilitate reading of the name.}

With the substance of the previous paragraph in mind, we could create the graph in Fig. 3.2 much more deliberately with the statements

```matlab
>> 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 )
```

\footnote{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.}
Here, we have used the command `figure` to create the figure object, the command `axes` to create an axes object within the figure object, and the command `line` to place the graph itself within the axes. Further, we have exploited several MATLAB 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 handles MATLAB 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

```matlab
>> hf = figure
>> ha = axes
>> hl = line( x, y )
>> htitle = title( 'Hyperbolic Cosine Function' )
```

for example, create the various objects with default properties and store the handles for these objects in the indicated variables. Once the handles are available, statements like

```matlab
>> get( hf )
>> get( ha )
>> get( hl )
>> get( htitle )
```

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

```matlab
>> set( hf )
>> set( ha )
>> set( hl )
>> set( htitle )
```

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

```matlab
>> set( ha, 'FontSize', 14, 'XLim', [-4.0,4.0], 'YLim', [0.0,12.0], ...
    'XGrid', 'on', 'YGrid', 'on' )
>> set( hl, 'Color', 'black', 'LineWidth', 4 )
```

allow manipulation of individual properties after the object has been created.

Sometimes, we will forget to save handles that we will subsequently need. MATLAB'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

```matlab
>> 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

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

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 `get` commands used shortly will reveal the default values of the several properties.
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 on the line created after the figure and axes objects have been created.

### 3.5 Saving and Retrieving the MATLAB Session

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

```
save example a b c
```

saves the variables `a`, `b`, and `c` in a file named `example.mat` in the current directory.\(^\text{17}\)\(^\text{18}\) If no variables are specified, then the command `save` saves all of the variables currently in the MATLAB workspace;\(^\text{19}\) if no filename is specified, the information is saved in the file `matlab.mat`.

Saved variables can be retrieved using the `load` command. To restore the file just saved, for example, we would use the statement\(^\text{20}\)

```
load example
```

which loads into the MATLAB workspace all of the variables saved in the file `example.mat` and gives each the name it had originally. Again, if no filename is specified, the information is loaded from the file `matlab.mat`.

In the first instance, the function `save` saves MATLAB variables in a compact, binary, MATLAB-specific, but machine-independent format. Sometimes we need to save MATLAB variables as ASCII text files so that other programs such as a text editor can read them. Actually, the `save` command can also save data in ASCII format if one or another switch is added to the command. For example, the statements

```
save example.asc -ascii a
save example.dou -ascii -double a
```

will save the values of the variable `a` in single-precision and double-precision ASCII format, respectively. Consecutive values on a line are separated by spaces, though adding the switch `-tabs` will instruct MATLAB to separate values with tab characters instead. With these switches, there is no default file type, so we will have to specify one (and must avoid `.mat`, since choosing it will lead to confusion for the `load` command). Note, also, that saving more than one variable at a time in ASCII format will render the resulting file confusing to the `load` command. Except for these provisos, the statements

```
load example.asc
load example.dou
```

will reload the data into MATLAB. With this approach, however, the variable saved will be recreated with the name of the file—here `example`—rather than its original name.

At a lower level of coding, MATLAB 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

\(^{17}\) The current directory is also referred to as the default directory.

\(^{18}\) See Section 3.18 for a fuller discussion of this directory and of means to change it.

\(^{19}\) In Windows, selecting ‘Save Workspace As . . .’ from the File menu brings up a browser in which a file name can be specified. With this route, however, the entire workspace must be saved.

\(^{20}\) If the current directory is unclear or ill-defined, specification of the full path to the file may be needed.
3.5. SAVING AND RETRIEVING THE MATLAB SESSION

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

in which—note—there are two spaces separating each %5d from the next, will write the values in that variable as ASCII text. First, the command fopen opens a file in the current directory for writing ('w'),\(^{21}\) 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 be 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 programs in C. In the string actually used above, the specification %5d indicates that the value should be output as an integer (d-format) right-justified in a field 5 characters wide.\(^{22}\) 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 MATLAB gives to the file. Reading the file into MATLAB at some later date cannot, however, be done with the load command. Instead, the more elaborate process with the statements

\[
\begin{align*}
&\text{>> } b = \text{zeros( 5, 5 );} & &\text{Create array of proper size.} \\
&\text{>> id = fopen( 'textexample.dat', 'r' );} & &\text{Open file for reading.} \\
&\text{>> b = fscanf( id, '%5d', [5,5] );} & &\text{Read data as a 5 x 5 array into } b. \\
&\text{>> status = fclose( id );} & &\text{Close file.}
\end{align*}
\]

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.\(^{23}\)

One warning about the structure of the file is necessary. For the example in this section, we assumed that we had a 5 x 5 matrix. In writing the elements of that matrix into the specified file, MATLAB will step down the first column of the matrix, which contains five values, writing those values one at a time into the first row—i.e., line—in the file. Then, MATLAB will step down the second column, writing its five values in turn into the second row (line) of the file. The process continues until the last column has been placed in the last row (line) 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 MATLAB manuals.

A final means to write MATLAB output to a file uses the command diary, which saves to a file everything that is printed in the MATLAB command window. The one argument to this

---

\(^{21}\) 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 more general formats %d, %f, and %e give MATLAB 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 3.7 and in the MATLAB manuals.

\(^{23}\) 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.
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

\[
\begin{align*}
&>> \text{dairy matrix_diary.dat} & \text{Open diary file in current directory.} \\
&>> a = [1,3,0; 3,7,2; 0,2,4] & \text{Create and display two matrices.} \\
& a = & \begin{bmatrix} 1 & 3 & 0 \\ 3 & 7 & 2 \\ 0 & 2 & 4 \end{bmatrix} \\
&>> b = [1,2,3; 4,5,6; 7,8,9] & \text{Multiply matrices element by element.} \\
& b = & \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \\
&>> c = a .* b & \text{Multiply matrices as matrices.} \\
& c = & \begin{bmatrix} 1 & 6 & 0 \\ 12 & 35 & 12 \\ 0 & 16 & 36 \end{bmatrix} \\
&>> d = a * b & \text{Close the file.} \\
& d = & \begin{bmatrix} 13 & 17 & 21 \\ 45 & 57 & 69 \\ 36 & 42 & 48 \end{bmatrix}
\end{align*}
\]

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

### 3.6 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 MATLAB. The simplest
is the \texttt{for} loop. Storing the squares of the integers from 1 to 10 in a vector, for example, is readily
accomplished with the statements\textsuperscript{24,25}

\[
\begin{align*}
&>> \text{for } i = 1:10 \\
& \quad x(i) = i^2; \\
& \text{end} \\
&>> x \\
& x = & \begin{bmatrix} 1 & 4 & 9 & 16 & 25 & 36 & 49 & 64 & 81 & 100 \end{bmatrix}
\end{align*}
\]

The \texttt{for/end} statement (which logically is a \textit{single} statement) steps through all values—integers
from 1 through 10, inclusive—of the loop index \texttt{i}, calculating their squares and storing each in
the proper element of \texttt{x}\textsuperscript{26}, and the statement \texttt{x} simply displays the end result.\textsuperscript{27} We make two further observations: (1) If the vector \texttt{x} already exists in MATLAB’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

\textsuperscript{24}Note that use of \texttt{i} for the loop index will override MATLAB’s default understanding of \texttt{i} as $\sqrt{-1}$. The default
meaning will be restored by the statement \texttt{clear i}.

\textsuperscript{25}The coding here illustrated requires MATLAB to adjust the size of the array \texttt{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 \texttt{x = 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{26}Without the semicolon after the one statement in the \texttt{for} loop, we would have been presented at \textit{each} step with
a display of the values stored in \texttt{x} at that step.

\textsuperscript{27}Alternatively, of course, we might have invoked the statements \texttt{y=1:10} followed by \texttt{x=y.^2}, but these statements
would not have illustrated the explicit construction of a loop.
all other elements untouched; a safer procedure would be to invoke the statement `clear x` before executing the loop; (2) MATLAB will also correctly execute the single-line presentation

```matlab
>> for i = 1:10 x(i) = i^2; end
```

though the semicolon would be replaced with a comma if we had 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

```matlab
>> clear x
>> for i = 1:2:10 x(i) = i^2; end
>> x
x = 1 0 9 0 25 0 49 0 81
```

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 MATLAB 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 output is to be suppressed). Taking the second option, we would achieve the desired objective with the statements

```matlab
>> clear x
>> for i = 1:2:10 x(i) = i^2; [i, x(i)], end
ans = 1 1
ans = 3 9
ans = 5 25
ans = 7 49
ans = 9 81
```

In addition to the `for` loop, in which we must know ahead of time how many times the loop is to be executed, MATLAB 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

```matlab
>> 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, the loop will, in fact, terminate after the tenth pass.

In constructing the `while` loop in the previous paragraph, we used MATLAB’s operator `<` (less than) to express our first logical condition. MATLAB, of course, possesses the standard six such operators, specifically `<` (less than), `>` (greater than), `==` (equal to), `~=` (not equal to), `<=` (less than or equal to), and `>=` (greater than or equal to). In addition, the operator `~` toggles 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 MATLAB operators & (logical and) and | (logical or) and the logical function xor (logical exclusive or).

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

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

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, MATLAB 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

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

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

```matlab
if a > 0
    b = a
else
    b = -a
end
```

will set b equal to the absolute value of a (though the function 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.

### 3.7 Reading Data from a File

Frequently, we need to read data into MATLAB 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 MATLAB are described in this section. Before the data can be read, MATLAB needs information about the type of data and the size and number of arrays in the file. If the file has any headers, MATLAB needs to be told where to start reading the real data.

We illustrate the process with the file 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 load as described in Section 3.5. 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. Commas or tabs can also be used to separate values in a line.
are the quantities of the initial material, the intermediate material, and the final (stable) material, respectively, at that time. We read the data from this file with the statements

```matlab
>> id = fopen('radio.dat', 'r');       % Open the file for reading, assign id.
>> ln = fgetl(id);                    % Read first line from file.
>> quantity = fscanf( id, '%f', [4,201] );   % Read data. The 4 x 201 array quantity
                                           % will be created and expanded as the data
                                           % are read.
>> status = fclose(id);                % Close file.
>> quantity = transpose(quantity);     % Transpose to 201 x 4 array.
```

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

```matlab
>> plot( quantity(:,1), quantity(:,2), 'Color', 'black', 'LineWidth', 4 )
>> set( gca, 'FontSize', 14, 'Box', 'off' )
>> hold on
>> plot( quantity(:,1), quantity(:,3), 'Color', 'black', 'LineWidth', 4 )
>> plot( quantity(:,1), quantity(:,4), 'Color', 'black', 'LineWidth', 4 )
>> text( 2.5, 850.0, 'A', 'FontSize', 16 )
>> text( 2.5, 300.0, 'B', 'FontSize', 16 )
>> text( 40.0, 800.0, 'C', 'FontSize', 16 )
>> hold off
```

The first of these statements plots the quantity of the initial material (quantity(:,2), the second column of the array) versus time (quantity(:,1), the first column of the array). The second statement enlarges the labels on the tic marks on the axes and turns off the box that otherwise would enclose the plot area. Then, since we executed the statement hold on, the fourth and fifth statements overlay graphs of the quantities of the second and third materials versus time. Finally, the statements invoking the command text, in which locations are expressed in the coordinates of the graph, place the labels A, B, and C at appropriate points on the graph.

Several new features of MATLAB have been introduced in this example:

- 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 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 assigned the value 16 to the property FontSize to override a somewhat smaller default size.

- The axes property Box, which controls whether a bounding box is drawn around the plotting area.

- 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.

---

29 Do not fail to copy this file into your directory and examine its structure with a text editor.

30 Remember that MATLAB will read the rows of the file into the columns of the array.

31 In MATLAB’s vocabulary, the command text places a text object within the already created axes object. This command will also return a handle to the text object it creates.

32 The command hold functions by manipulating both the axes property NextPlot and the figure property NextPlot.
CHAPTER 3. INTRODUCTION TO MATLAB

Figure 3.3: Simulation of radioactive decay.

The resulting plot is shown in Fig. 3.3.

Before leaving this example, we comment on additional features of the command plot. Among other useful features (see the MATLAB 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

\[
\begin{align*}
&\text{>> } t = \text{quantity}( :, 1 ); \\
&\text{>> } abc = \text{quantity}( :, 2:4 ); \\
&\text{>> } \text{plot( } t, abc \text{ )}
\end{align*}
\]

In this form, MATLAB will draw each new graph with a different color, cycling through the colors stored in the axes property ColorOrder.\(^{33}\)

3.8 On-line Help

MATLAB provides several ways for users to request help while in a MATLAB session. As we have already seen, we can obtain information about the current status of the MATLAB 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 help will produce a list of topics about which further information can be displayed with the statement

\[
\text{help topic}
\]

e.g., help elfun, help specfun, and help graph2d. These statements, in turn, will produce lists of actual commands about which detailed information can be obtained with the statement

\(^{33}\)Color is more fully discussed in Section 3.17.5.
3.9. M-FILES

Table 3.2: The MATLAB script plotradio.m. Comments introduced by a percent 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 MATLAB.

```matlab
id = fopen('radio.dat', 'r'); % Open the file, 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 = transpose(quantity); % Transpose array

plot(quantity(:,1), quantity(:,2), ... % Plot A vs t
     'Color', 'black', 'LineWidth', 4)
hold on
plot(quantity(:,1), quantity(:,3), ... % Add B vs t
     'Color', 'black', 'LineWidth', 4)
plot(quantity(:,1), quantity(:,4), ... % Add C vs t
     'Color', 'black', 'LineWidth', 4)
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
```

help command

e.g., help plot, help sin, and help meshgrid. More conveniently, the command helpwin will bring up the MATLAB Help window, in which mouse clicks can be used to select from a menu of available topics. Finally, the command helpdesk will open up a browser and give access to an extensive MATLAB manual, and the command demo will bring up a window that will allow browsing through a wide assortment of interactive demonstrations of particular commands and features. Many of these windows can also be opened by selecting an item from the Help menu in the GUI.

3.9 M-Files

Any command file called by MATLAB 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. MATLAB makes available many built-in functions. We have already introduced several (plot, fopen, inv, zeros, fix, sin, cosh, sqrt, ...). Others include functions for drawing error bars, invoking filters, selecting color tables, and producing animated displays.

MATLAB’s built-in functions can be supplemented with user-written scripts and functions. The script is the simpler of these options. Using an available text editor, 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 3.2. Then, after this file has been stored with the name plotradio.m (say) in a directory that is in MATLAB’s search path (see Section 3.18 and the Local Guide), the single statement\(^{34}\)

```matlab
>> plotradio
```

\(^{34}\)When no file type is specified, the file type '*.m' is assumed by default.
### Table 3.3: The MATLAB function `luplot.m`

```matlab
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, 'LineWidth', 2); % Plot graph
end
```

will instruct MATLAB to execute each of these statements in turn and ultimately produce the graph we have already shown in Fig. 3.3. 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.

We can also define customized procedures and functions. The reserved word `function` on the first non-blank line of the file informs MATLAB 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 3.3 define a function that plots the function `funct`, evaluating it at `N+1` equally spaced points spanning the interval `start ≤ x ≤ stop`. While the final `end` statement is optional in MATLAB, best coding practice includes it anyway. 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 MATLAB’s search path, either the statement

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

which uses the newer syntax, or the statement

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

which uses an older syntax, will generate a graph of the cosine function \( \cos(x) \) over the interval \( 0 ≤ x ≤ 10 \), dividing the interval into 200 segments (i.e., plotting 201 points uniformly distributed over the interval). Note carefully the single quotation marks enclosing the name of the function in the older form.

What if, however, we wanted to use `luplot` to graph a function that is not built into MATLAB? In that case we would first write a function M-file defining the function of interest. For example, suppose we wanted a graph of the Lorentz lineshape defined by

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

(3.4)

35Since 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 `[ ] =`.

36See the last paragraph of this section for a discussion of the significance of the comments before the first executable statement in this file.

37See Section 3.18.
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);
end

for particular values of $a$, $b$, and $x_0$. We would begin by writing a function M-file defining the
Lorentz lineshape, perhaps using the coding listed in Table 3.4. Then, after the file lineshape.m
has been stored somewhere in MATLAB’s search path,\textsuperscript{38} we simply use a statement like

\begin{verbatim}
>> l unwrap( @lineshape, 200, -10.0, 10.0 )
\end{verbatim}

to call l unwrap. In a second or two, the graph of the Lorentz lineshape shown in Fig. 3.4 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 l unwrap
is written, however, we cannot simply add these parameters as arguments to lineshape. Were we
to adopt that approach, we would want to set the parameters $a$, $b$, and $x_0$ at MATLAB’s command

\textsuperscript{38}See Section 3.18.
Table 3.5: The MATLAB function `lineshape1.m`.

```matlab
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);
end
```

level and have them known not within `luplot` itself but within a function that `luplot` calls. The statement `feval` in `luplot` does not include the passing of parameters to `lineshape`. The easiest means to achieve this passing of a parameter from MATLAB’s command level into a function that is not directly called from the command level involves using `global` variables.\(^\text{39}\) 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 MATLAB’s automatic isolation of variables in functions from variables of the same name in other functions, rendering the variables known in any function that declares them `global`—and `only` in those functions. To this end, we might rewrite `lineshape` to be as listed in Table 3.5 and store it in the file `lineshape1.m`. Here, the statement `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 `global` and assigned values at command level.\(^\text{40}\) Then, we invoke `luplot` with `this` function using the statements

\[
\begin{align*}
>> & \text{global A B X0} \quad \text{Declare global variables.} \\
>> & A = 1.0, B = 1.0, X0 = 2.0 \quad \text{Set parameters.} \\
>> & \text{luplot( @lineshape1, 200, -10.0, 10.0)} \quad \text{Graph function.}
\end{align*}
\]

and produce again the graph shown in Fig. 3.4. 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 MATLAB’s capacity for self-documentation. We have again introduced comments with percent signs. In function M-files, comment lines that begin with a percent sign and `precede` the first executable line in the file have special significance. The statement

\[\text{lookfor CharacterString}\]

searches the first comment line (i.e., the first line that starts with a percent sign) of every M-file that it finds and displays every such line containing the specified character string. Further, the statement

\[\text{help FileName}\]

\(^\text{39}\)Alternatively, we could refer to the MATLAB manuals and exploit the function `varargin`, which supports the use of a variable number of arguments, as a means to pass parameters without using global variables. This route is significantly more involved than the more straightforward—though risky—use of global variables.

\(^\text{40}\)We are following a recommendation—not a mandate—of The MathWorks in using 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.
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.*

### 3.10 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, finding principal axes and moments of inertia, ...). The MATLAB 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

```matlab
>> A = [ 1,3,0; 3,7,2; 0,2,4 ]
A =
1 3 0
3 7 2
0 2 4
>> evals = eig(A)
evals = 
-0.3893
3.4469
8.9425
```

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

```matlab
>> [evecs,evals] = eig(A)
evecs = 
 0.8912 -0.3107  0.3305
-0.4127 -0.2534  0.8749
 0.1881  0.9161  0.3540
evals = 
-0.3893  0 0
    0 3.4469 0
    0 0 8.9425
```

With this statement, each column of the array `evecs` is one of the *normalized* eigenvectors of the matrix `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

$$Ax = \lambda x$$

(3.5)

where $\lambda$ is an eigenvalue and $x$ is the corresponding eigenvector. For the first eigenvalue and eigenvector, we find that

```matlab
>> lambda = evals(1,1);
>> x = evecs(:,1);
>> [A * x, lambda * x]
an = 
-0.3470 -0.3470
 0.1607  0.1607
-0.0732 -0.0732
```

Extract first eigenvalue.

Extract first eigenvector.

Construct two-column array with left side of Eq. (3.5) as first column and right side as second column.

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. (3.5), we simply evaluate

```matlab
>> lhs = A * evecs
```

```
lhs =
-0.3470  -1.0708   2.9552
  0.1607  -0.8734   7.8238
-0.0732   3.1578   3.1659
```

To obtain the right hand side, however, we recognize that the diagonal character of the matrix of eigenvalues means that we must evaluate the product in a seemingly unnatural order, specifically

```matlab
>> rhs = evecs * evals
```

```
rhs =
-0.3470  -1.0708   2.9552
  0.1607  -0.8734   7.8238
-0.0732   3.1578   3.1659
```

to have each column in the result be the product of the corresponding column in `evecs` times the proper eigenvalue in `evals`. 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

```matlab
>> rhs - lhs
```

```
ans = 1.0e-14 *
 -0.0222  -0.0444   0.0888
  0.0028  -0.1998   0.0888
  0.0167      0   0.0888
```

where the premultiplier $10^{-14}$ applies to all values. Evidently, the two matrices, whose elements are all on the order of 1, differ by rather less than $10^{-14}$, surely equal within the roundoff errors endemic to (double precision) computer calculations.\footnote{Note that, because these results are influenced by roundoff, the actual values received may vary from platform to platform and from version to version of MATLAB. All values, however, are essentially zero.}

MATLAB’s command `eig` is quite general. The matrix on which it works must, of course, be square, but it need not be symmetric or real. Further, with a second argument, e.g. `eig(A,B)`, the command can find eigenvalues and eigenvectors for the more general problem $Ax = \lambda Bx$.

### 3.11 Graphing Scalar Functions of One Variable

The most common—and simplest—graph is a two-dimensional plot of a function of one variable, i.e., a graph of dependent variable versus independent variable. As we have already seen in Section 3.3.5, 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 the properties `FontSize`, `LineWidth`, `Color`, `XLim`, `YLim`, `XGrid`, and `YGrid`. Further, graphs can be embellished by invoking additional functions, of which we have already met `title`, `xlabel`, `ylabel`, `hold`, and `text`. Other properties and functions will be introduced as the need for them arises.
3.11.1 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

$$B(z) = \frac{1}{(1 + z^2)^{3/2}}$$  \hfill (3.6)

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 MATLAB 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 )
>> set( gca, 'FontSize', 14 )
>> title( 'Magnetic Field on z axis', ...
       'FontSize', 20 )
>> xlabel( 'Dimensionless Position', ...
        'FontSize', 16 )
>> ylabel( 'Magnetic Field', ...
      'FontSize', 16)
```

where the properties Color, LineWidth, and FontSize and the functions title, xlabel, ylabel, set, and gca have the effects described in Sections 3.3.5 and 3.4. This sequence of statements will produce the graph shown in Fig. 3.5.

3.11.2 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

```matlab
>> dx = 8.0/100.0;
>> x = [-4.0 : dx : 4.0 ];
>> y1 = sin(x);
>> y2 = cos(x);
>> plot( x, y1, 'Color', 'black', ...
      'LineWidth', 4 ), plot( x, y2, 'Color', 'red', ...
      'LineWidth', 4 )
>> set( gca, 'FontSize', 14 )
>> title( 'Sine and Cosine on z axis', ...
       'FontSize', 20 )
>> xlabel( 'Dimensionless Position', ...
        'FontSize', 16 )
>> ylabel( 'Magnetic Field', ...
      'FontSize', 16)
```
Figure 3.6: Undamped and damped sine waves.

```
>> x = [ 0.0 : 0.1 : 10.0 ];
>> ys = sin(x); yc = cos(x);
>> plot( x, [ ys; yc ] );
```

One useful feature of this approach is that each new superimposed graph will be displayed in a different color. Further, the handle returned by this use of `plot` will be a two-component vector whose first component is the handle for the first line drawn and whose second component is the handle for the second line drawn.

MATLAB’s command `hold`, which we first met in Section 3.7, provides an alternative means for plotting more than one graph on a single set of axes. When overplotting, a different line style may be used to draw each of the graphs. The plot of undamped and damped sine waves on the same axes shown in Fig. 3.6 is produced with the statements

```
>> dx = 20.0/100.0;
>> x = [ 0.0 : dx : 20.0 ];
>> sine = sin(x);
>> dampsine = exp(-x/10.0) .* sin(x);
>> plot( x, sine, 'Color', 'black', 'LineWidth', 2 )
>> title( 'Damped and Undamped Sine Waves', 'FontSize', 20 )
>> set( gca, 'FontSize', 14 )
>> hold on
>> plot( x, dampsine, 'Color', 'black', 'LineWidth', 2, ...  
       'LineStyle', '-.' )
>> hold off
```

The property `LineStyle` specifies a line style other than the default (a solid line). Table 3.6 enumerates the available line styles. Note that the codes for these line styles can be combined with the abbreviations for colors in Table 3.1 so that the second statement involving `plot` above could be written

```
>> plot( x, dampsine, 'k-.', 'LineWidth', 2)
```

42 The sequence of colors can be printed on the screen by executing the statement `get(gca, 'ColorOrder' )` when the desired set of axes is the current set. The first three colors are blue, green, and red. Additional information about the use of color is presented in Section 3.17.5.
3.11. GRAPHING SCALAR FUNCTIONS OF ONE VARIABLE

Table 3.6: Available line styles for use with the property LineStyle.

<table>
<thead>
<tr>
<th>Value for linestyle</th>
<th>Line Style Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>Solid (default)</td>
</tr>
<tr>
<td>:</td>
<td>Dotted</td>
</tr>
<tr>
<td>-</td>
<td>Dashed</td>
</tr>
<tr>
<td>-.</td>
<td>Dash Dot</td>
</tr>
</tbody>
</table>

Figure 3.7: Polar plot of the cardioid

3.11.3 Polar Plots and Custom Axes

MATLAB is also capable of producing polar plots and of using different axis styles instead of the standard box style. For example, to graph the cardioid defined in polar coordinates by the equation

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

and shown in Fig. 3.7, we might execute the statements

```
>> dt = 2.0*pi/100.0;
>> theta = [ 0 : dt : 2*pi ];
>> a = 3.0;
>> r = a*(1.0 - cos(theta));
>> hp = polar( theta, r )
>> set( gca, 'FontSize', 14 )
>> set( hp, 'LineWidth', 4 )
>> title( 'The Cardioid', 'FontSize', 20 )
```

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 handle.
Enlarge tic labels.
Make line heavier.
Title display.

Here, in the command `polar`, the azimuthal coordinate—in radians—comes first and the radial coordinate comes second. Note also that polar graphs are automatically displayed with equal increments on both axes; that feature need no be explicitly stipulated.
3.11.4 Multiple Plots On a Page

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

- m specifying the number of plot rows in the Figure window,
- n specifying the number of plot columns in the Figure window, and
- 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 wish to plot a 2 × 2 array of graphs, so we will use the statement subplot(2,2,?). In total, we produce the plot shown in Fig. 3.8 with the statements

```
>> 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', 4 )
>> title( 'Sine', 'FontSize', 20 )
>> subplot(2,2,2)
>> plot( z, w, 'Color', 'black', 'LineWidth', 4 )
>> title( 'Cosine', 'FontSize', 20 )
>> subplot(2,2,3)
>> plot( z, x, 'Color', 'black', 'LineWidth', 4 )
>> title( 'Hyperbolic Sine', 'FontSize', 20 )
>> subplot(2,2,4)
>> plot( z, y, 'Color', 'black', 'LineWidth', 4 )
>> title( 'Hyperbolic Cosine', 'FontSize', 20 )
>> subplot(111)
```

Note (1) that we have used semicolons to separate individual statements placed on a single line and (2) that subplot is reset to specify one plot per page at the end of this sequence of statements.43

3.11.5 Plotting Experimental Data

As a final example of two-dimensional graphing, we describe MATLAB’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. MATLAB will also allow us to set the range of the axes and annotate the graph with a legend. This example uses data from an experiment on an RC high-pass filter. Suppose the file rcd data.dat containing the data was created with a text editor and stored in the current directory or in a directory in MATLAB’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 MATLAB workspace with the statements

```
43The special form subplot(111) without commas sets a flag that causes the next graphics command to refresh the entire Figure window. This form also does not set a handle, so providing a variable for a returned value constitutes an error.
```
Figure 3.8: Graphs of sine, cosine, hyperbolic sine, and hyperbolic cosine on the interval $-4 \leq x \leq 4$.

Next, we plot the data using the statements

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

Open data file.
Read data.
Close file.
Transpose array.

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. Using the coding in Table 3.7, the property `Marker` specifies which of several symbols is to be placed at each data point. Absence of a code for a line style will suppress the line altogether. 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. 3.9.

A graph of experimental data, however, is not complete without error bars. MATLAB’s command `errorbar` is up to the task, though we need to be careful to suppress its default connecting of the points with straight lines. Supposing a symmetric uncertainty in the vertical coordinate of $\pm 5\%$, we add the error bars with the statements

```matlab
>> semilogx( data(:,1), data(:,2), 'Color', 'black', ...
            'Marker', 'diamond', 'LineStyle', 'none' )
>> axis( [ 10.0, 10000.0, 0.0, 1.1 ] )
>> title( 'Experimental Data for High Pass Filter', 'FontSize', 20 )
>> xlabel( 'Frequency', 'FontSize', 16 )
>> ylabel( 'Gain (Vout/Vin)', 'FontSize', 16 )
```

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.

Using the symbols in the first column of Table 3.7, we could abbreviate the above use of the command `semilogx` with the statement

```matlab
>> semilogx( data(:,1), data(:,2), 'kd' )
```

in which we have also used a color code from Table 3.1.

As described in the MATLAB manuals, the command `errorbar` has numerous options, including the possibility of drawing asymmetric limits and of drawing horizontal as well as vertical limits.
Table 3.7: Available symbols for use with the property \texttt{Marker}.

<table>
<thead>
<tr>
<th>Value of \texttt{Marker}</th>
<th>Symbol Drawn</th>
<th>Value of \texttt{Marker}</th>
<th>Symbol Drawn</th>
</tr>
</thead>
<tbody>
<tr>
<td>.</td>
<td>point</td>
<td>v</td>
<td>triangle (down)</td>
</tr>
<tr>
<td>o</td>
<td>circle</td>
<td>'</td>
<td>triangle (up)</td>
</tr>
<tr>
<td>x</td>
<td>times sign</td>
<td>&lt;</td>
<td>triangle (left)</td>
</tr>
<tr>
<td>+</td>
<td>plus sign</td>
<td>&gt;</td>
<td>triangle (right)</td>
</tr>
<tr>
<td>*</td>
<td>star</td>
<td>p</td>
<td>pentagram</td>
</tr>
<tr>
<td>s</td>
<td>square</td>
<td>h</td>
<td>hexagram</td>
</tr>
<tr>
<td>d</td>
<td>diamond</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ unc = \texttt{data(:,2)} \times 0.05; \]

\[ \texttt{hold on} \]

\[ \texttt{errorbar( data(:,1), data(:,2), unc, '.k');} \]

Compute uncertainties at 5\%. (This vector might alternatively be read from a properly expanded input file.)

Hold existing plot.

Draw error bars at each point. In the specifier \texttt{'.k'}, the dot replaces automatic connection of the data points with straight line segments and the \texttt{k} stipulates drawing the error bars with the color black.

Note that, once the logarithmic scale has been established with a previous invocation of \texttt{semilogx}, the logarithmic scaling will not be altered by subsequent plotting statements.

A graph of experimental data should also include a plot of predicted theoretical results. The equation for the theoretical gain of this circuit is

\[
G = \frac{V_{\text{out}}}{V_{\text{in}}} = \frac{1}{\sqrt{1 + \left(\frac{f_0}{f}\right)^2}} \quad \text{where } f_0 \text{ is defined as } f_0 = \frac{1}{2\pi RC} \quad (3.8)
\]

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 \texttt{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 \texttt{temp}, i.e.

\[
f = 10^{\text{temp}} = \left( e^{2.3026} \right)^{\text{temp}} = e^{2.3026 \times \text{temp}} \quad (3.9)
\]

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.\footnote{See also the function \texttt{logspace} described in the MATLAB manuals.}

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

\[ \texttt{r = 1.5e4;} \quad \texttt{Set value of } R. \]
\[ \texttt{c = 0.0442e-6;} \quad \texttt{Set value of } C. \]
\[ \texttt{f0 = 1/(2.0*pi*r*c);} \quad \texttt{Calculate implied reference frequency.} \]
\[ \texttt{temp = [ 1.0 : 0.12 : 4.0 ];} \quad \texttt{Create vector with values from 1.0 to 4.0.} \]
\[ \texttt{f = exp(2.3026 * temp);} \quad \texttt{Create frequencies from } 10^1 \text{ to } 10^4. \]
\[ \texttt{gtheory = 1.0./sqrt(1.0 + f0^2/f.^2);} \quad \texttt{Calculate theoretical curve.} \]
\[ \texttt{plot( f, gtheory, 'Color', 'black', ...} \quad \texttt{Overplot theoretical curve.} \]
\[ \quad 'LineStyle', '--', 'LineWidth', 2 )} \]
Here, we have used the values $R = 15 \, \text{k}\Omega$ and $C = 0.0442 \, \mu\text{F}$ and then overplotted the theoretical curve with a dashed linestyle.

Finally, to annotate the graph with a legend reminding us of the meaning of each of its pieces, we would use the statements

```matlab
>> plot( [500.0, 900.0], [0.3, 0.3], 'LineStyle', '--', 'LineWidth', 2, 'Color', 'black' )
>> text( 1000.0, 0.30, 'Theoretical', 'FontSize', 16 )
>> plot( [700.0], [0.21], 'Marker', 'diamond', 'Color', 'black', 'LineStyle', 'none' )
>> text( 1000.0, 0.20, 'Experimental', 'FontSize', 16 )
>> hold off
```

The first of these statements uses `plot`—remember, we have already set the logarithmic scale on the horizontal axis—to draw a line from data coordinate $(500.0, 0.3)$ to the data coordinate $(900.0, 0.3)$. The second then uses `text` to place the label “Theoretical” near to the line, with the label starting at the point $(1000.0, 0.3)$. The remaining statements place a diamond plotting symbol and label it “Experimental”.

The statements listed above and used to produce Fig. 3.9 have been compiled into the file `plotrc.m`, which is available where all other files mentioned in this chapter are found.

### 3.12 Making Hard Copy

#### 3.12.1 ... of Text

To print mathematical results from MATLAB when MATLAB is being run on an X-window device or in Windows, we can use the cutting and pasting capabilities of the X-window system/clipboard to copy the information from the window running MATLAB to the input window of the text editor in use. Alternatively (and necessarily on devices without cut/paste capability), the command `diary` described in Section 3.5 will create an ASCII file containing everything written to the screen in a MATLAB session, which file can then be edited to extract only the portions wanted.
3.12.2 ... 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 a PostScript description of the plot.\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.} Then, we must send that file to the printer. The second of these steps is a task for the operating system and is described in the Local Guide. The first step, however, is a task for MATLAB. To create a file containing an encapsulated level 2 black and white PostScript description of the contents of the current Figure window, issue the simple MATLAB statement

\[\text{print -deps2 FileName}\]

If you wish the file to contain an encapsulate PostScript (Level 2, color) description of the display, issue instead the statement

\[\text{print -depbc2 FileName}\]

Other device specifications are listed in the MATLAB manuals and in the message displayed in response to the statement \text{help print}.\footnote{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.}

Note that some features of a graph as it appears on the screen may not be exactly replicated in the output produced by a PostScript file. In particular the effect of the properties \texttt{FontSize} and \texttt{LineWidth} in the plot command may change when a new output device is selected. Thus, when these (and perhaps other) properties are specified, some experimentation may be necessary to produce a \textit{printed} graph that exactly reflects what may have been displayed on the screen.

3.13 Graphing Scalar Functions of Two Variables

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

\[z(x, y) = \sin(2\pi x) \sin(3\pi y)\] \hspace{1cm} (3.10)

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

\[I(\xi, \eta) = \left(\frac{\sin \xi}{\xi}\right)^2 \left(\frac{\sin \eta}{\eta}\right)^2\] \hspace{1cm} (3.11)

which is related to the intensity in the diffraction pattern produced by a square aperture. In graphing a function of a \textit{single} variable, we used as input to \texttt{plot} the values of the function at a selected set of (regularly or irregularly spaced) values spanning the desired range of that variable. \textit{Here}, we need as input to any graphing routine the values of the function at a regular,\footnote{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.} two-dimensional \textit{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 \((x, xi, \ldots)\) containing the first coordinate \((x, \xi, \ldots)\) of all the points in the grid and the other \((y, eta, \ldots)\) 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 MATLAB can process arrays element by element, a single statement, which would have the forms
3.13. GRAPHING SCALAR FUNCTIONS OF TWO VARIABLES

\[ z = \sin(2.0\pi x) \times \sin(3.0\pi y); \]
\[ I = (\sin(xi)/xi)^2 \times (\sin(eta)/eta)^2; \]

for the functions in Eqs. (3.10) and (3.11), 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 \texttt{mesh}, \texttt{contour}, and \texttt{surf}.

3.13.1 A Preliminary: The Function \texttt{meshgrid}

In MATLAB, the two-dimensional analog of the statement \(x = [\text{start : dx : stop}]\) is provided by the command \texttt{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\(^{50}\)

\[ \text{>> x} \]
\[ x = \]
\[ -3 -1 1 3 \]
\[ -3 -1 1 3 \]
\[ -3 -1 1 3 \]
\[ -3 -1 1 3 \]
\[ \text{>> y} \]
\[ y = -1.0000 -1.0000 -1.0000 -1.0000 \]
\[ -0.5000 -0.5000 -0.5000 -0.5000 \]
\[ 0 0 0 0 \]
\[ 0.5000 0.5000 0.5000 0.5000 \]
\[ 1.0000 1.0000 1.0000 1.0000 \]

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

\[ \text{>> } z = x.^2 + y.^2 \]
\[ z = \]
\[ 10.0000 2.0000 2.0000 10.0000 \]
\[ 9.2500 1.2500 1.2500 9.2500 \]
\[ 9.0000 1.0000 1.0000 9.0000 \]
\[ 9.2500 1.2500 1.2500 9.2500 \]
\[ 10.0000 2.0000 2.0000 10.0000 \]

\(^{50}\)Note that the division of the \(x\) interval \([-3.0,3.0]\) into \textit{three segments} results in \textit{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 \textit{four segments} results in \textit{five} values \((-1, -0.5, 0, 0.5, 1)\).
Consider the dot before the symbol $^\wedge$ 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 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

```matlab
>> 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( -3.0*pi:dxi:3.0*pi, -3.0*pi:dxi:3.0*pi );  
>> I = (sin(xi)./xi).^2 .* (sin(eta)./eta).^2; 
```

To create more refined arrays $z$ and $I$ for the two examples of Eqs. (3.10) and (3.11). (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 $I$.)

### 3.13.2 Surface Plots: The Function `mesh`

The simplest MATLAB function for displaying two-dimensional arrays is `mesh`, and the statements

```matlab
>> mesh( z )
>> mesh( I )
```

will generate a three-dimensional wire-mesh surface conveying the altitudes in the two-dimensional arrays $z$ and $I$. The specification of additional arguments and properties, however, will create a more refined display. Executing the more elaborate statements

```matlab
>> mesh( x, y, z, 'EdgeColor','black' )
>> title('[2,3] Mode', 'FontSize', 20 )
>> xlabel('x', 'FontSize', 16 )
>> ylabel('y', 'FontSize', 16 )
>> zlabel('z', 'FontSize', 16 )
>> mesh( xi, eta, I, 'EdgeColor','black' )
>> title('Diffraction (Square Aperture)', 'FontSize', 20 )
>> xlabel('\xi', 'FontSize', 16 )
>> ylabel('\eta', 'FontSize', 16 )
>> zlabel('I', 'FontSize', 16 )
```

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 `mesh` (or `xi`; `y` or `eta`) tell MATLAB 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

---

51 In the second of the statements involving `xlabel` and `ylabel`, we have invoked MATLAB’s capability to use special symbols. In particular, the Greek letters can be printed by preceding the name of the letter with a backslash. Here, for example, `\xi` produces $\xi$ and `\eta` produces $\eta$. Those familiar with the way $\TeX$ and $\LaTeX$ produce these special symbols will recognize the constructions used by MATLAB. These features will be more fully described in Section 3.17.1.
3.13. GRAPHING SCALAR FUNCTIONS OF TWO VARIABLES

-10.0 ≤ ξ, η ≤ 10.0). The third argument provides the values of the function to be represented as a surface, and the stipulation of the property `EdgeColor` sets the color used to draw the edges of each patch in the display, overriding a coloring that links edge color to the value of the function. The more complete graphs produced by these last statements are shown in Figs. 3.10 and 3.11. Note that, in the perspective shown, the x and ξ axes run positively to the east northeast, the y and η axes run positively to the west northwest, and the z and I axes run positively to the north (vertical).

An auxiliary command `view` gives access to the two optional properties that control the angle from which the surface is viewed. The `azimuth` specifies a rotation, in degrees, from the negative y axis towards the positive x axis in the xy plane and is, by default, −37.5°; the `elevation` specifies the angle of rise out of the xy plane towards the positive z axis and is, by default 30°. The azimuth and elevation, together with the assumption that the viewer is quite distant from the surface, locate the viewpoint from which the surface will have the displayed appearance. To change the azimuth to +20° while leaving the elevation at +30°, we would execute the statement

```matlab
>> view([20, 30])
```

when the display we wish to reorient is located in the current `Figure` window. Figures 3.12 and 3.13 show the resulting rotated views for comparison with Fig. 3.10 and Fig. 3.11.

Examination of Figs. 3.11 and 3.13 suggests that the pattern might have weak peaks (bright spots) 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 0.1])
```

preserves the automatically determined ranges on the ξ and η axes but forces the I axis to be confined to the range 0.0 ≤ I ≤ 0.1 (and the central peak simply goes off scale). The resulting display, shown in Fig. 3.14, reveals the secondary peaks in the diffraction pattern and even hints at some of those not on the main axes.
3.13.3 Contour Plots: The Function \textit{contour}

The function \textit{contour}, whose general syntax is similar to that of the command \textit{mesh}, draws contour lines for two-dimensional arrays. Using all of the defaults, we would produce a contour map for the [2,3]-mode of the membrane with the statement

\begin{verbatim}
>> contour(z)
\end{verbatim}

Exploiting a few properties, however, we could produce a more meaningful contour map with the statements\textsuperscript{52}

\begin{verbatim}
>> contour( x, y, z, 20, 'k' )
\end{verbatim}

where, as with \textit{mesh}, the first and second arguments define the scales for the two axes while the third argument provides the data for generating the display. The fourth argument—here 20—specifies the number of levels to be drawn, overriding the default, and the brief argument \textit{'k'} replaces the default coloring of contour lines and makes them all black. Further, we could replace the specification of the \textit{number} of contour lines with a vector giving the \textit{values} at which contour lines are to be drawn as, for example, in the statement\textsuperscript{53}

\begin{verbatim}
>> contour( x, y, z, [-0.8,-0.6,-0.4,-0.2,0.2,0.4,0.6,0.8], 'k' );
\end{verbatim}

Even better, we could draw \textit{and label} the positive contour lines with the statements

\begin{verbatim}
>> [ Cpos, Hpos ] = contour( x, y, z, [0.2,0.4,0.6,0.8], 'k' );
>> clabel( Cpos, Hpos )
\end{verbatim}

\textsuperscript{52}Note that the order of the arguments in this statement is important. In particular, the argument \textit{'k'}—which \textit{cannot} be replaced with the more extended version \textit{'Color'}, \textit{'black'}—must be \textit{last}.

\textsuperscript{53}Again, the argument \textit{'k'} must be \textit{last}. 
and then add labeled negative contour lines with the statements

\[
\begin{align*}
&\text{ >> hold on} \\
&\text{ >> [ Cneg, Hneg ] = contour( x, y, z, [-0.8,-0.6,-0.4,-0.2], 'k--' );} \\
&\text{ >> clabel( Cneg, Hneg )}
\end{align*}
\]

Here, we have exploited the capability of \texttt{contour} to return a \textit{contour matrix} (\texttt{Cpos} and \texttt{Cneg}) and a vector of \textit{handles} (\texttt{Hpos} and \texttt{Hneg}), both of which are needed as arguments for the auxiliary command \texttt{clabel}, which places labels on each line. Finally, we could recognize that we are dealing with a \textit{square} membrane and use the statement

\[
\begin{align*}
&\text{ >> axis square} \\
&\text{ >> hold off}
\end{align*}
\]

to establish a 1:1 aspect ratio for the display.\textsuperscript{54} The end result of these several operations is the display shown in Fig. 3.15. Further details not only for the command \texttt{contour} but also for the related command \texttt{contourf}, which draws \textit{filled} contour maps, can be found in the MATLAB manuals.

3.13.4 Shaded Surfaces: The Function \texttt{surf}

The function \texttt{mesh} displays a two-dimensional array using a “wire-mesh” technique in which data points are connected to adjacent points using lines. MATLAB can also generate an alternate display in which a smooth but shaded surface conveys the values in a two-dimensional array. To produce these representations of our two examples, we might accept all defaults and execute the statements

\[
\begin{align*}
&\text{ >> surf( z )} \\
&\text{ >> surf( I )}
\end{align*}
\]

\textsuperscript{54}A fuller discussion of the command \texttt{axis} is presented in Section 3.17.4.
With a bit more care, we can produce a more sophisticated display. For example, we could use the statements

```
>> surf( x, y, z )
>> shading interp
>> title( '[3,2] Mode', 'FontSize', 20 )
>> xlabel( 'x', 'FontSize', 16 )
>> ylabel( 'y', 'FontSize', 16 )
>> zlabel( 'z', 'FontSize', 16 )
>> colormap gray
```

to

- produce an initial display (with `surf`).
- change the shading from the default `faceted`, which keeps the same color in each patch of the surface as defined by the lines shown by `mesh` (and shows those lines), to `interp`, which interpolates the coloring to produce a smoother gradation from region to region (and does not show the lines),\(^{55}\)
- label the display and its axes, and
- change the colormap to a gray scale (see Section 3.17.5).

These statements—and nearly identical ones for the function \( I(\xi, \eta) \)—generate the displays shown in Fig. 3.16 and 3.17, though the printed versions using a gray scale pale by comparison with the color plots on the screen.

\(^{55}\)The remaining option is `flat`, which is the same as `faceted` except it also does not show the lines.
Figure 3.14: A rescaled surface representation of the function $I$ in Eq. (3.11).

Figure 3.15: Contour plot of the function $z$ in Eq. (3.10). This view shows the membrane from above, where solid lines represent regions in which the membrane is displaced towards the viewer and dashed lines represent regions in which the membrane is displaced away from the viewer.
3.13.5 Features of the *Figure* Window

Exploration of the nature of a surface in two dimensions is greatly facilitated by several features of the *Figure* window. Three icons in the toolbar in this window enable zooming in, zooming out, and rotating the display. By default, each of these features is off. Clicking ML on any one of these icons, however, will toggle the feature from its current (on or off) to the other state. If, for example, *zoom in* is enabled (on), clicking ML with the cursor at a particular point in the display will move the viewpoint one step towards the identified point while clicking MR will move the viewpoint one step away from the identified point. Additional clicks of ML or MR will move the viewpoint additional steps in the same direction. With *zoom out* enabled, the behavior is similar, except that the roles of ML and MR are reversed. *Warning*: Be careful not to click MR or ML repeatedly too quickly; the processor may not be able to keep up with the changes thereby requested.

Finally, if rotate 3D is enabled, pressing and holding any mouse button when the cursor is somewhere on the display picks up the display and permits its reorientation in real time by moving the mouse. When the mouse button is released, the display is fixed at its new perspective. Note that, when the mouse button is being held, a pop-up message in the lower left corner of the *Figure* window displays the azimuth and elevation of the current orientation—values that could well be useful in the subsequent construction of a statement exploiting the command *view* to produce the identical view from the command line.

3.13.6 The Functions *meshc*, *surfc*, and *surfl*

A more complex way to represent a two-dimensional array would use 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

---

56 The features can also be selected from the *Tools* menu in the *Figure* window.
3.13. GRAPHING SCALAR FUNCTIONS OF TWO VARIABLES

Figure 3.17: Shaded surface representation of the function \( I \) in Eq. (3.11).

\begin{align*}
\text{Diffraction (Square Aperture)}
\end{align*}

\begin{verbatim}
>> meshc( x, y, z )
>> meshc( xi, eta, I )
>> surfc( x, y, z )
>> surfc( xi, eta, I )
>> surf1( x, y, z )
>> surf1( xi, eta, I )
\end{verbatim}

will produce these displays for each of our sample functions. The details for using \texttt{meshc}, \texttt{surfc}, and \texttt{surf1} are fully described in the MATLAB manuals.

3.13.7 Functions of Two Variables in Polar Coordinates

MATLAB also makes available the function \texttt{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

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

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

\begin{verbatim}
>> z = (1 - r.^2).* exp(-r).*cos(phi);
>> mesh( x, y, z )
>> contour( x, y, z, 20 )
>> surf( x, y, z )
\end{verbatim}
Figure 3.18: Mesh representation of the function $z$ in Eq. (3.12).

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

Figure 3.19: Contour map of the function $z$ in Eq. (3.12).

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

$$z(r, \theta) = (1 - r^2) e^{-r} \cos \theta$$  \hfill (3.12)

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. 3.18 and Fig. 3.19.

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 MATLAB manuals.
3.14 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

\[
p(x, y, z) = \frac{1}{2\pi (27)^3} e^{-2\rho/3} \left(3 \frac{z^2}{\rho^2} - 1\right)^2 = \frac{1}{2\pi (27)^3} e^{-2\rho/3} (9z^4 - 6z^2\rho^2 + \rho^4) \tag{3.13}
\]

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

\[
\rho = \sqrt{x^2 + y^2 + z^2} \tag{3.14}
\]

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 MATLAB 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 \texttt{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. (3.13) is created using the statements

\[
\begin{align*}
&\text{Inc} = 20.0/32.0; & \text{Calculate increment.} \\
&[x, y, z] = \text{meshgrid}(-10.0 : \text{Inc} : 10.0, \\
&\quad \quad \quad \quad \quad \quad \quad \quad -10.0 : \text{Inc} : 10.0, \\
&\quad \quad \quad \quad \quad \quad \quad \quad -10.0 : \text{Inc} : 10.0); & \text{Calculate grid.} \\
&\text{Pfactor} = 1.0 / (2.0*\text{Pi}^2*27.0^3); & \text{Calculate premultiplier.} \\
&\text{Zs} = z.^2; & \text{Calculate } z^2. \\
&\text{Rhos} = x.^2 + y.^2 + \text{Zs}; & \text{Calculate } \rho^2. \\
&\text{Rho} = \text{sqrt(rhos)}; & \text{Calculate } \rho. \\
&p = \text{Pfactor} * \exp(-2.0*\text{Rho}/3.0) .* \ldots & \text{Calculate } p(x, y, z). \\
&\quad (9.0*\text{Zs}.*2 - 6.0*\text{Zs}.*\text{Rhos} + \text{Rhos}.*2);
\end{align*}
\]

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.

3.14.1 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\(^{57}\)

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

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{>> px = zeros(33,33);} & \quad \text{Preallocate storage for array.} \\
\text{>> for i = 1:33} & \quad \text{Create a floating point array.} \\
\text{for j = 1:33} & \quad \text{Extract data in plane parallel to \( yz \) plane.} \\
\text{\hspace{0.5cm} \text{px(i,j)} = p(15,i,j);} & \\
\text{end} & \\
\text{end} & \\
\text{>> sc = [-10.0 : inc : 10.0]}; & \quad \text{Create vector for scaling axes.} \\
\text{>> mesh( sc, sc, px, 'EdgeColor', 'black')}; & \quad \text{Generate and label mesh plot.} \\
\text{>> xlabel('x','FontSize', 14)} & \\
\text{>> ylabel('y','FontSize', 14)} & \\
\text{>> zlabel('z','FontSize', 14)} & \\
\text{>> contour( sc, sc, px, 20, 'k')}; & \quad \text{Generate and label contour plot.} \\
\text{>> xlabel('x','FontSize', 14)} & \\
\text{>> ylabel('y','FontSize', 14)} & \\
\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 \texttt{mesh} and \texttt{contour} can be \textit{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. 3.20 and Fig. 3.21.

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. MATLAB’s command \texttt{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 \texttt{slice} is

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

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

\[
\begin{align*}
\text{>> axis square} \\
\text{>> slice( } x, y, z, p, 0.0, 0.0, 0.0 \text{ );} \\
\text{>> shading interp} \\
\text{>> set( gca, 'Visible', 'off' )}
\end{align*}
\]

\(^{57}\)We must extract two-dimensional arrays from all three arguments, since these statements will not behave properly with three-dimensional arrays in any position.
Figure 3.20: Mesh representation of the function $p$ in Eq. (3.13) in the plane $x = x_0 + 15 \times dx = -10.0 + 15 \times 20.0/32.0 = -0.625$.

Figure 3.21: Contour map of the function $p$ in Eq. (3.13) in the plane $x = x_0 + 15 \times dx = -10.0 + 15 \times 20.0/32.0 = -0.625$.

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

```matlab
>> 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 $x = -7.5, 0.0, 7.5$, one parallel to the $zx$ plane at
CHAPTER 3. INTRODUCTION TO MATLAB

$y = 0$, and one parallel to the $xy$ plane at $z = 0$ and add a colorbar to facilitate interpreting the color code. The controls made available in the toolbar of the Figure window—see section 3.13.5—may be especially useful in helping to make the features of these displays understandable.

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' );
```

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

```
>> clf;
>> 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
```

Figure 3.22 shows the display produced by these statements, though the black and white rendition pales by comparison with the color rendition on the screen.

### 3.14.2 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
The first statement—*isosurface*—searches the array $p$ to generate in $\text{poly}$ 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 second statement—*patch*—draws the isosurface in a *Figure* window, though it doesn’t generate a suitable perspective in that window. Thus, we need the third statement—*axis*—to establish a proper transformation of the data. The remaining statements make adjustments in the coloring and lighting of the surface. Some trial and error was necessary to determine suitable values. The resulting output is shown in Fig. 3.23.

### 3.15 Graphing Vector Fields

Graphical display of *vector* functions of two or three variables is more complicated than display of scalar functions, but MATLAB 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. 3.24 when the electric field has only a $z$ component, the magnetic field has only $x$ and $y$ components, and none of

---

58 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 $\text{max}$ and $\text{min}$, 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 $\text{max}(\text{max}(\text{max}(p)))$. 

---
these components depends on $z$. In this case, the only non-zero components of the electric field $E$ and the magnetic field $H$ are given at time $t = 0$ by

$$
E_z = \cos \kappa_x b \tau \sin n\pi \eta,
$$

$$
H_x = \sin \kappa_x b \tau \cos n\pi \eta,
$$

$$
H_y = -\frac{\kappa_x b}{n\pi} \cos \kappa_x b \tau \sin n\pi \eta
$$

(3.15)

\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.}
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, $\overline{x} = x/b$ and $\overline{y} = y/b$ are dimensionless coordinates within the guide, and
\[
\left( \frac{\kappa_x b}{n\pi} \right)^2 = \left( \frac{\omega b}{n\pi c} \right)^2 - 1
\]  
(3.16)

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. (3.16) 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
\[
E_z = \cos 2\pi \overline{x} \sin 2\pi \overline{y} \\
H_x = \sin 2\pi \overline{x} \cos 2\pi \overline{y} \\
H_y = -\cos 2\pi \overline{x} \sin 2\pi \overline{y}
\]  
(3.17)

where $\overline{x}$—the coordinate along the guide can range over any values—we choose $0 \leq \overline{x} \leq 1$—but, to be inside the guide, $\overline{y}$ is confined to the region $0 \leq \overline{y} \leq 1$.

Each component of these fields can now be represented by a two-dimensional array. The $H$ field, which has two non-zero components, then is translated into two such arrays; the $E$ field, which has only one non-zero component, requires only one such array. These arrays are readily created with the MATLAB statements

```matlab
>> [ x, y ] = meshgrid( 0.0:0.04:1.0, 0.0:0.04:1.0 );
>> xp = 2.0*pi*x; \n| \n| yp = 2.0*pi*y;
>> Ez = cos(xp) .* sin(yp);
>> Hx = sin(xp) .* cos(yp);
>> Hy = -cos(xp) .* sin(yp);
```

3.15.1 The Function quiver

The MATLAB 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

```
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 $H$ in MATLAB’s memory can be displayed in this form by executing the statement

```
>> quiver( x, y, Hx, Hy, 'k' )
>> xlabel( 'x', 'FontSize', 16 )
>> ylabel( 'y', 'FontSize', 16 )
```

The resulting display is shown in Fig. 3.25.

\[60\] From here on in this section, we drop the overbars.
CHAPTER 3. INTRODUCTION TO MATLAB

Figure 3.25: Magnetic field of Eq. (3.17) 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.

3.15.2 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

```matlab
>> hold on
>> contour( x, y, Ez, [0.1,0.3,0.5,0.7,0.9], 'k' )
>> contour( x, y, Ez, [-0.9,-0.7,-0.5,-0.3,-0.1], 'k-.')
>> hold off
```

will add contour lines to Fig. 3.25, producing the display shown in Fig. 3.26. 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.

3.15.3 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, MATLAB provides the function `quiver3` (see the MATLAB 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.
3.16 Animation

Along with all of the image processing and graphing capabilities, MATLAB can also display a sequence of images in rapid succession to produce an animated display. 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.\(^{61}\) The statements creating a three-dimensional byte array containing the 41 images in the animation are

```matlab
>> id = fopen('ldrum.dat', 'r');  
>> x = [ 0.0 : 1.0/32.0 : 1.0 ];  
>> for j=1:41  
      disp = fscanf( id, '%f', [33,33] );  
      mesh( x, x, disp', 'EdgeColor', 'black' )  
      axis( [ 0.0 1.0 0.0 1.0 -0.25 0.25 ] )  
      M(j) = getframe;  
>> end  
>> status = fclose( id );
```

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

\(^{61}\)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.

\(^{62}\)Here, we have used the shorthand symbol ‘\(^\dagger\)’ to transpose \(\texttt{disp}\) rather than the longer function \texttt{transpose}. Strictly, the symbol ‘\(^\dagger\)’ 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 \textit{without complex conjugation}, we would use \texttt{transpose} or the special symbol ‘\(^\dagger\)'.
Several options for displaying the images can be invoked. All use the command `movie` but with different arguments. The statement

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

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

\[ \text{>> movie( } M, 15 \text{ );} \]

exploits an optional (integer) argument to effect a 15-fold display of the full sequence of images,\(^63\) and the statement

\[ \text{>> movie( } M, [ 15 1 3 5 7 9 ] \text{ );} \]

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

\[ \text{>> movie( } M, [ 1 23 ] \text{ );} \]

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

\[ \text{>> movie( } M, [15 1 3 5 7 9], 2 \text{ );} \]

exploits a further argument that controls the speed of the display—here set to 2 frames per second. As with all MATLAB commands, we can use \( \text{(CONTROL-C)} \) to interrupt a movie before it is finished, though the MATLAB 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. MATLAB 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.\(^64\) 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

\(^63\)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.

\(^64\)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.
3.17. ADVANCED GRAPHING FEATURES

```
>> [ 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
```

Create array describing initial shape.
Create values of time.
For each of 41 images:
Calculate data for image.
Create display.
Standardize axes.
Read display as bit map.

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.

### 3.17 Advanced Graphing Features

In this section, we describe some of MATLAB’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.

#### 3.17.1 Fonts

Among the most useful of MATLAB’s auxiliary features is a capability to choose characters from a wide variety of standard fonts (including Roman, Italic, Greek, and 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 MATLAB’s capabilities but are in no way a substitute for careful reading of the MATLAB manuals.

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

- **FontName**, which can assume any of the values `AvantGarde`, `Bookman`, `Courier`, `Helvetica`, `Helvetica-Narrow`, `NewCenturySchlbk`, `Palatino`, `Symbol`, `Times-Roman`, `ZapfChancery`, and `ZapfDingbats`. MATLAB substitutes `Courier` for any name not in this list.
- **FontSize**, which is expressed in 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.
- **FontAngle**, which is normal by default but can be changed to italic or oblique.

By default, these properties are set in the axes object, and can be changed by appropriate specifications either in the command `axes` itself, e.g.,

```
axes( ..., 'FontName', 'Palatino', 'FontSize', 16, ... )
```

or, after axes have been created, by analogous specifications with the command `set`, e.g.,

```
set( gca, 'FontName', 'Palatino', 'FontSize', 16, ... )
```
Table 3.8: Some codes for use in character strings in MATLAB. 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.

<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>\vartheta</td>
<td>$\vartheta$</td>
<td>\sim</td>
<td>$\sim$</td>
<td>\geq</td>
<td>$\geq$</td>
</tr>
<tr>
<td>\varphi</td>
<td>$\varphi$</td>
<td>\leq</td>
<td>$\leq$</td>
<td>\propto</td>
<td>$\propto$</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>\angle</td>
<td>$\angle$</td>
<td>\rightarrow</td>
<td>$\rightarrow$</td>
<td>\neg</td>
<td>$\neg$</td>
</tr>
<tr>
<td>\approx</td>
<td>$\approx$</td>
<td>\uparrow</td>
<td>$\uparrow$</td>
<td>\nabla</td>
<td>$\nabla$</td>
</tr>
<tr>
<td>\cdotp</td>
<td>$\cdot$</td>
<td>\downarrow</td>
<td>$\downarrow$</td>
<td>\ldots</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>\times</td>
<td>$\times$</td>
<td>\circlearrowleft</td>
<td>$\circlearrowleft$</td>
<td>\pm</td>
<td>$\pm$</td>
</tr>
<tr>
<td>\surd</td>
<td>$\sqrt{\ }$</td>
<td>\rightarrow</td>
<td>$\rightarrow$</td>
<td>\neg</td>
<td>$\neg$</td>
</tr>
</tbody>
</table>

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

```matlab
hdl = get( gca, 'Title' )
set( hdl, 'FontName', 'Symbol', ... )
```

get the handle for the text object containing—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 entire object. 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

```matlab
title( 'This {\fontname{symbol}\fontsize{10}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 the word ‘is’ will be presented in the symbol font (and come out as $\vartheta$) at 10 pt. Here, braces play two roles: (1) they enclose the arguments of the commands \fontname and \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.}

Several other Greek letters and mathematical symbols are available with commands that mimic those used in \LaTeX. All lower case Greek letters except omicron (which looks like a Roman o (‘oh’)) can be inserted with a command like \alpha (letter name in lower case preceded by a backslash); all 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 \Gamma (letter name with initial letter in upper case preceded by a backslash). Other available symbols are listed in Table 3.8. Thus, for example, the statement

```matlab
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 ‘$^\cdot$’; subscripts are created with the character ‘$_\cdot$’.
3.17 ADVANCED GRAPHING FEATURES

Figure 3.27: A more elegant graph of the $z$ axis magnetic field of a current carrying loop.

Two further points: (1) While in-string selection of \fontName and \FontSize can be accomplished with the commands \fontname and \fontsize, parallel in-string specification of \fontWeight and \fontAngle is not available. Instead, we must use one of \bf, \sl, \it, and \rm to produce bold, slanted, italic, and Roman styles, respectively, and—to limit the scope of the indicated change—we enclose the characters in braces, e.g., \{\bfBoldWord(s)\}. (2) The characters \_, \^, \, \{, and \} 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 \_, \^, \, \{, and \}, respectively, in the quoted string.

In Section 3.13.2, 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. 3.5 by (1) changing the font used for the various titles and (2) replacing the title on the $y$ axis with the equation itself. The statements

\begin{verbatim}
>> 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( '{\itB}(\{itz\})=1/(1+\{itz\}^2)^{3/2}', 'FontSize', 16 )
\end{verbatim}

will produce the revised graph of the magnetic field shown in Fig. 3.27.

3.17.2 Space Curves

Sometimes it is desirable to view the trajectory of a particle in three-dimensional space. MATLAB 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{equation}
\begin{align*}
x &= \cos t & \quad y &= \sin t & \quad z &= \alpha t
\end{align*}
\end{equation}
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 )
```

will produce the graph in Fig. 3.28.

### 3.17.3 Using Multiple Windows

In the simple circumstances so far introduced, we have accepted MATLAB’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\(^{66}\)

```matlab
hdl = figure
hdl = figure( 'Position', ...
             [400,200,500,300] )
```

```matlab
hdl = figure( 'MenuBar', 'none' )
hdl = figure( 'Name', 'New Window' )
figure(hdl)
```

\(^{66}\)Normally, of course, each of these statements would be terminated with a semicolon to suppress explicit output.
3.17. ADVANCED GRAPHING FEATURES

Get handle of current window.

Clear current window but leave window in place.

List current values of properties of window hdl.

List properties and possible values for window hdl.

Delete 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 brings it to the fore and makes it the current window.

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

```
>> 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.

3.17.4 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 tight**, which turns off MATLAB's effort to find "nice" numbers for the ranges on the axes and forces the ranges to be whatever is represented in the actual plotted data.

- **axis auto**, which restores the default auto-scaling on the axes.

- **axis normal**, which restores the default size of the plotting area and restores the default auto-scaling on the axes.

Restoring all defaults without recreating the axes object will frequently require both **axis auto** and **axis normal**, though the single statement

```
axis auto normal
```

will accomplish both objectives.

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 **equal**, **square**, and **tight**, we execute the statements
Figure 3.29: The effect of different axis configurations.

\[ t = \left[ 0.0 : 2.0\pi/100.0 : 2.0\pi \right]; \]
\[ y = \sin(t); \]
\[ x1 = \cos(t) + 1.25; \quad x2 = x1 - 2.5; \]
\[ \text{subplot}(2,2,1) \]
\[ \text{plot}(x1,y), \text{hold on}, \text{plot}(x2,y), \text{title('defaults')} \]
\[ \text{subplot}(2,2,2) \]
\[ \text{plot}(x1,y), \text{hold on}, \text{plot}(x2,y), \text{axis equal}, \text{title('equal')} \]
\[ \text{subplot}(2,2,3) \]
\[ \text{plot}(x1,y), \text{hold on}, \text{plot}(x2,y), \text{axis square}, \text{title('square')} \]
\[ \text{subplot}(2,2,4) \]
\[ \text{plot}(x1,y), \text{hold on}, \text{plot}(x2,y), \text{axis tight}, \text{title('tight')} \]

The resulting display is shown in Fig. 3.29. Notice that only the specification \texttt{axis equal} creates circles that appear as circles.

### 3.17.5 Working with Color

As discussed briefly in Section 3.3.5, the elemental stipulation of a color in MATLAB is provided by a three-component vector that defines the RGB components of the color. Each component ranges from 0 to 1. Thus, for example, the vector \[ [0.25 \hspace{1em} 0.40 \hspace{1em} 1.00] \] conveys a color that has red intensity 0.25, green intensity 0.40, and blue intensity 1.00. The RGB designations for several standard colors were presented in Table 3.1.

Colors used by MATLAB are stored in the system color table, which contains two parts. \textit{Fixed} colors—however many there are—are stored as the property \texttt{FixedColors} of a figure object. The value of this property is a three-column matrix, each row of which conveys the RGB components of a single color. Creation of a figure object, either directly with \texttt{figure} or indirectly with other commands like \texttt{plot} or \texttt{mesh}, initializes a default table of fixed colors containing just three colors. Adding an axes object as a child of the figure object appends seven more colors—the colors through which MATLAB cycles by default when more than one line is drawn in a single invocation of \texttt{plot}.

When a statement like

```matlab
\texttt{text( 'Title', 'Color', [0.4,0.2,0.9] )}
```
is executed, the specified color is added to the table of fixed colors. Further, the statement

```matlab
get( gcf, 'FixedColors' )
```

will display the current fixed color list.

The second part of the system color table is called the *colormap* and does not interact with the fixed color list. The colormap is stored as the property `ColorMap` of a figure object. The value of this property is a three-column matrix, each row of which conveys the RGB components of a single color. By default, a 64-color colormap named `jet` is assigned to the property `ColorMap` when a new `Figure` window is created. To change the colormap, we first need to create a three-column matrix `M` containing the RGB vectors for all the colors we wish to define in the order in which we want them to be associated with numeric values of increasing magnitude. Then, we simply execute the statement

```matlab
colormap( M );
```

to replace the current colormap with the new one. To return to the default, we would execute the statement

```matlab
colormap jet(64)
```

where `jet` is a function that creates a particular colormap, here containing 64 colors. Besides `jet`, MATLAB makes available numerous other functions to create various colormaps, some of them being `hot`, `cool`, `spring`, `summer`, `autumn`, `winter`, `gray`, `copper`, and `bone`. The available shorthand names for a number of standard color maps are listed in the message that is printed with the statement `help graph3d`. Further, the statement `help ColorMapName` will bring up a description of each colormap. Finally, since displays produced with `mesh` and `surf` use the current colormap, recreating Fig. 3.16, for example, and then executing statements like

```matlab
colormap spring
colormap hot
```

will display some characteristics of each selected map in turn.

### 3.18 Miscellaneous Occasionally Useful Tidbits

In this section, we enumerate features of MATLAB that are used so infrequently that they are easily forgotten between uses.

#### 3.18.1 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.

---

67 The specific number of colors needs to be included to allow for the possibility that `M` may have contained a different number of colors. The function `jet` by itself returns a colormap having the same number of colors as the current map.
3.18.2 MATLAB’s Current Directory

If MATLAB is launched with a command typed in the Shell window to the operating system, 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 MATLAB is launched. If, on the other hand, MATLAB 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 3.18.4). In the GUI, the identity of the current directory is conveyed in a text-entry box labeled ‘Current Folder’ in the toolbar at the top of the window and also a text entry box in the panel labeled ‘Current Folder’.

The directory can be edited by changing the entry in either of those boxes by browsing in the associated drop-down menu, by explicitly typing the desired directory in either box, or by selecting from the folders displayed in the panel labeled ‘Current Folder’. 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.

3.18.3 MATLAB’s Search Path

Whenever an explicit path is not specified for a file that MATLAB must read (either data file or M-file), the program looks first in the current directory, and then in a sequence of directories defined by the search path. The initial search path is defined by default or by a MATLAB-startup file that is read when MATLAB is launched, and it is site-specific (see the Local Guide). MATLAB keeps track of the current search path in the system variable **path**, which can be printed with the statement **path**. Typically, the search path will begin with a directory /Documents/MATLAB in the user’s home directory and continue with several directories containing files from the initial MATLAB installation.

The path can be changed within a specific session with MATLAB by assigning a new value to this variable. The statements

```
addpath 'Directory1ToAdd', 'Directory2ToAdd', ... -begin
addpath 'Directory1ToAdd', 'Directory2ToAdd', ... -end
```

will place the directories *Directory1ToAdd, Directory2ToAdd, ...* at the beginning (**-begin**) or end (**-end**) of the current search path while the statement

```
path( path, 'DirectoryToAdd' )
```

will add the new directory at the end of the current path and the statement

```
path( 'DirectoryToAdd', path )
```

will add the new directory at the beginning of the current path (or move it to the beginning of the end if the directory is already in the search path). The **path** function can be used to add only one directory at a time to the search path; **addpath** can add several directories in a single statement.

The statement **rmpath** *Directory*—see MATLAB manuals—will remove an existing directory from the current path.

---

68This panel will be displayed in the default appearance of the GUI.
3.18.4 Customizing MATLAB

The configuration of MATLAB can be customized in two ways. First, the system manager can define a system-wide startup file—a batch file that is executed immediately after MATLAB is started. This file is named `matlabrc.m` and is stored in the directory `$MATLABHEAD/toolbox/local`, but—except on single-user systems—can be changed only by the system manager. The last action in this file, however, involves executing the file `startup.m`, should it happen to exist in a directory in the MATLAB search path, presumably the first directory in that search path. Thus, individual users can alter the “standard” starting environment with commands placed in this special file.

3.18.5 Restoring MATLAB’s Initial State

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

```
>> clear -all
>> startup
```

3.18.6 The Command `format`

Beyond the formats noted in Section 3.3.1, MATLAB also accepts the statements

```
format compact
format loose
```

the first of which causes MATLAB to use less space in output it directs to the screen and the second of which restores the default.

3.18.7 Suppressing Axes

We have already seen that axes and their labels can be suppressed altogether by executing the simple statement

```
set( gca, 'Visible', 'off' )
```

which turns off the axes and sets the background color in the plotting area to be the same as the border color in the figure object of which the axes object is a child. Alternatively, we might “hide” the axes by setting their color to that of the background in the axes object and also setting the border color in the figure object to that color. The statements

```
bck = get( gca, 'Color' );
set( gca, 'XColor', bck, 'YColor', bck, 'ZColor', bck )
```

find the background color in the current axes and then set the border color in the current figure and the colors of the three axes (and their labels) to that color. These statements will *not* change titles and labels created by `title`, `xlabel`, `ylabel`, `zlabel`, or `text`.

---

69 You may have to invoke the command `path` to discover what that directory is. See Section 3.18.3.
3.18.8 Positioning Axes

The position of the $x$ axis is controlled by the property `XAxisLocation`, which has the value `bottom` by default but can be changed to `top`; the position of the $y$ axis is controlled by the property `YAxisLocation`, which has the value `left` by default but can be changed to `right`.

3.18.9 Distinguishing Graphs on a Single Set of Axes

In Section 3.7 and in Fig. 3.3, we placed three separate graphs on a single set of axes and distinguished graphs from one another by placing an appropriate label ($A$, $B$, and $C$) near to each graph. Suppose we once again execute the statements

```matlab
>> id = fopen('radio.dat', 'r');              % Open the file, assigning id.
>> ln = fgetl(id);                           % Read first line from file.
>> quantity = fscanf( id, '%f', [4,201] );    % Read data. The array `quantity` will be created and expanded as the data are read.
>> status = fclose(id);                      % Close file.
>> quantity = transpose(quantity);           % Transpose array.
```

to read in the data from which Fig. 3.3 was generated but then, for shorthand in what follows, we extract the first, second, third, and fourth columns of the array `quantity`, assigning a shorter name to each with the statements

```matlab
>> t = quantity(:,1); A = quantity(:,2); B = quantity(:,3); C = quantity(:,4);
```

In this abbreviated notation, the simple statement

```matlab
>> plot( t, [A, B, C], 'LineWidth', 4 );
```

produces a graph similar to Fig. 3.3, except that the line representing $A$ is displayed in blue, the line representing $B$ is displayed in green, and the line representing $C$ is displayed in red. Thus, if a graph containing several lines is produced by including a list of dependent variables in a single invocation of `plot`, the colors used will cycle through a sequence determined by the axes property `ColorOrder`—a 3 column by $m$ row matrix of RGB values. The current value of that property can be displayed with the statement

```matlab
>> get( gca, 'ColorOrder' )
```

By default, the displayed matrix will have seven rows and will specify a cycling through the colors blue, green, red, cyan, magenta, yellow, and gray—though not all will be as intense as those illustrated in Table 3.1. The user, however, can control the colors in the cycle by setting the `figure` property `DefaultAxesColorOrder`, as described in the MATLAB manuals.

If the graph is ultimately to be printed in black and white, cycling line styles rather than colors provides a more appropriate means to distinguish graphs on the same set of axes. By default, all lines are drawn solid. The user can control that feature by setting the `figure` property `DefaultAxesLineStyleOrder`. For example, the statements

```matlab
>> set( gcf, 'DefaultAxesColorOrder', [0 0 0] );
>> set( gcf, 'DefaultAxesLineStyleOrder', '-|--|-.' );
>> plot( t, [A, B, C], 'LineWidth', 4 );
```
3.18. MISCELLANEOUS OCCASIONALLY USEFUL TIDBITS

will produce a graph in which \( A \) is shown with a solid line, \( B \) with a dashed line, and \( C \) with a dash-dot line. Had we more than three dependent variables to plot, the cycle solid, dashed, dash-dot would be repeated until all dependent variables had been plotted.

The properties `ColorOrder` and `LineStyleOrder` interact with one another. In general, `plot` will start with the first line style specified by `LineStyleOrder`, execute one cycle through the colors specified by `ColorOrder`, then select the second line style, cycle again (once) through the colors, .... In order for each new graph to take the next line style, we must be sure that `ColorOrder` contains only one color. The first statement in the code in the previous paragraph sets that one color to black.

3.18.10 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 MATLAB command `input` achieves this objective. Its syntax is illustrated in the statement

\[
\text{x = input('Enter the value of x: ')}
\]

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 \( x \). 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

\[
\text{pos = input('Enter } x \text{ and } y \text{ coordinates: ')}
\]

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 \( \text{pos} \) will be a vector.

3.18.11 Writing Data to the Screen

In MATLAB, 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 semicolon will not be enough. The code segments

\[
\begin{align*}
\text{disp('This is a message.'),} \\
\text{a = [0.0 10.0; 0.1 15.0; 0.2 18.0];} \\
\text{disp( ' t (s) x (cm)')}, \\
\text{disp( a )}, \\
\text{a11 = num2str(a(1,1), '¥4.1f'),} \\
\text{a12 = num2str(a(1,2), '¥5.1f'),} \\
\text{str = ['When } t \text{ = ', a11, ' s, x = ', a12, ' cm.']}, \\
\text{disp( str )}, \\
\text{fprintf( 1, 'When } t \text{ = %4.1f s, x = %5.1f cm.' )}
\end{align*}
\]

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.
3.18.12 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

```matlab
for i = 1:n
    if 2*fix( A(i)/2 ) == A(i)
        break;
    end
end
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.

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

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

encountered during execution will display the error message and return control immediately to whatever level called the procedure of function.

3.19 References

Those seeking additional information about MATLAB should

- Consult the on-line help messages that can be brought up by typing `help`, `help topic`, or `help command` at the MATLAB prompt or selecting ‘Product Help’ (or a similar phrase) from the HELP menu in the `MATLAB command` window.

- Google 'MATLAB manuals' and examine some of the links that emerge, including in particular
  - the link to MATLAB documentation, [www.mathworks.com/help/matlab](http://www.mathworks.com/help/matlab), which provides access to numerous documents provided by the vendor of MATLAB.
  - several links to documents produced by third-parties.

You may also find in the library associated with your computational facilities one or more of the manuals

- *Getting Started with MATLAB*, which—in less than 100 pages—illustrates the more commonly used capabilities of MATLAB.
• Using MATLAB, which provides a comprehensive description of all features of MATLAB, including getting started, manipulating the desktop, using functions, using the help system, working with files, using and writing M-files, and many other topics.

• Using MATLAB Graphics, which focuses on MATLAB’s capabilities for visualizing data graphically and supplements the volume Using MATLAB.


• Building GUIs with MATLAB, which describes how to create MATLAB applications that are controlled from GUIs.

• MATLAB Application Program Interface Guide, which—among other advanced features—describes how to call external programs written in FORTRAN or C from within MATLAB.

While the versions available to you may be outdated (and some may no longer be in print), most of the information in any that you can find is almost certainly still accurate.

3.20 Exercises

3.20.1 Writing MATLAB Statements

3.1. Write and test MATLAB 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.

3.2. Look up the function sort both in the on-line help and in the printed MATLAB 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.

3.3. Look up the function lu both in the on-line help and in the printed MATLAB manuals, then use the process described therein to solve the linear equations solved by other means in Section 3.3.4.

3.4. Describe and test a sequence of MATLAB statements that uses a for/end loop to evaluate the dot product of two $n$-component vectors $a$ and $b$. 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.

3.5. (a) Describe and test a sequence of MATLAB statements that uses a for/end loop to evaluate $\sum a_i$ when the values of $a_i$ are supplied as the elements of the vector $a$. 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 MATLAB 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 when it is applied to a vector. (d) Explore and describe the behavior of sum and cumsum when applied to a matrix rather than a vector.

3.6. Write and test a MATLAB function that accepts two three-component vectors as input and returns a three-component vector containing the cross product of the two input vectors.

3.7. 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.

3.20.2 Finding Eigenvalues and Eigenvectors

3.8. 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.

3.9. 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 $10 \times 10$ 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 $10 \times 10$ 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

```matlab
>> 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

```matlab
>> x = 0.0 : 1.0/11.0 : 1.0
>> 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 $50 \times 50$.

3.10. 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 $\langle nlm|eF\hat{z}|n'l'm'\rangle$, 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

\[
3ea_0F
\begin{pmatrix}
0 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 \\
-1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}
\]

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 $|3,2,2\rangle$, $|3,1,1\rangle$, $|3,2,1\rangle$, $|3,0,0\rangle$, $|3,1,0\rangle$, $|3,0,1\rangle$, $|3,0,2\rangle$, $|3,1,2\rangle$, and $|3,2,0\rangle$, then the matrix for the state $n = 3$ is...
3.20. EXERCISES

3.20.3 Graphing Scalar Functions of a Single Variable

3.11. Create and test a sequence of MATLAB statements that will produce a graph that is similar to Fig. 3.6 except that the axes are drawn along the lines $y = 0$ and $x = 0$.

3.12. (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 $c = 1.0$, (c) explore the way the function depends on the three parameters $a$, $b$, and $c$, and (d) write a paragraph describing that behavior in words.

3.13. 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 $z/a$.

3.14. 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.

3.15. 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

$$I(\delta) = \frac{1}{1 - 2r \cos \delta + r^2}$$

Obtain graphs of $I(\delta)$ versus $\delta$ for various values of $r$, and write a paragraph describing these graphs.

3.16. 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 $s$
CHAPTER 3. INTRODUCTION TO MATLAB

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

\[
m = \frac{m_0}{\sqrt{1 - \beta^2}} ; \quad p = \frac{m_0 v}{\sqrt{1 - \beta^2}} ; \quad K = \frac{m_0 c^2}{\sqrt{1 - \beta^2}} - m_0 c^2
\]

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.

3.18. 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

\[
T = \frac{4n^2}{4n^2 + (n^2 - 1)^2 \sin^2(\kappa d)} ; \quad R = \frac{(n^2 - 1)^2 \sin^2(\kappa d)}{4n^2 + (n^2 - 1)^2 \sin^2(\kappa d)}
\]

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.

3.19. 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 \( b \) \((b > 0)\) 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

\[
V(z) = \frac{\sigma}{2\epsilon_0} \left[ \sqrt{R^2 + (z - b/2)^2} - \sqrt{R^2 + (z + b/2)^2} \right]
\]

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.

3.20. Using data from an experiment you have performed, create a suitable ASCII text file, read the data into MATLAB, and produce a graph showing the data, error bars, and a theoretical curve. Label the graph completely.

3.21. 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 a MATLAB procedure to produce such a graph and, if you have suitable data available (or if you can invent some), demonstrate that your procedure works.

3.22. 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.

3.23. 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

\[
E(\theta) = \frac{q}{4\pi \epsilon_0 r^2} \frac{1 - \beta^2}{(1 + \beta^2 \sin^2 \theta)^{3/2}}
\]
3.24. 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.

3.20.4 Graphing Scalar Functions of Two Variables

3.25. The Planck radiation law gives the expression

\[
u(\lambda, T) = \frac{8\pi ch}{\lambda^5} \frac{1}{e^{\lambda kT} - 1}
\]

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 \( \lambda_0 = 550 \text{ nm} \).) With these changes, the expression to be plotted can be recast in the form

\[
u(\lambda, T) = \frac{1/\Lambda^5}{8\pi ch / \lambda_0^3} e^{\lambda/kT_0} - 1
\]

and the question now becomes one of plotting this quantity using the dimensionless variables \( \Lambda \) and \( \tau \).

3.26. 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.

3.27. 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

\[
B(z) = \frac{1}{2} B_0 \left( a^2 + b^2 \right)^{3/2} \left[ \frac{1}{\left(a^2 + (z + b)^2\right)^{3/2}} + \frac{1}{\left(a^2 + (z - b)^2\right)^{3/2}} \right]
\]

where \( B_0 \) is the magnetic field at the origin. Plot graphs showing \( B(z)/B_0 \) (a) 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.
3.28. In an LRC circuit of resonant frequency $\omega_0$, the current $I$ is given as a function of frequency $\omega$ by

$$I = \frac{I_0}{\sqrt{1 + Q^2 \left(\frac{\Omega}{\omega_0} - 1\right)^2}}$$

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.

3.29. A spherical potato of radius $a$ is taken from the refrigerator at $0 \, ^\circ\text{C}$ and placed in an oven at $u_0 = 200 \, ^\circ\text{C}$. The temperature $u(r, t)$ at a point a distance $r$ from the center of the potato at time $t$ is given by

$$\frac{u(r, t)}{u_0} = 1 - 2 \sum_{n=1}^{\infty} \frac{j_0(\beta_n r/a)}{\beta_n j_1(\beta_n a)} e^{-\kappa \beta_n^2 t/(ca^2)}$$

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 $175 \, ^\circ\text{C}$, i.e., to a value such that $u(0, t)/u_0 = 0.875$. **Hints:** (1) Note that

$$j_0(x) = \frac{\sin x}{x} \quad ; \quad j_1(x) = \frac{\sin x}{x^2} - \frac{\cos x}{x}$$

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 $a = 0.05 \, \text{m}$ 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.

3.30. 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 MATLAB, 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 $GmM/R$.

### 3.20.5 Graphing Scalar Functions of Three Variables

3.31. Following the pattern illustrated in Section 3.14, explore at least one of the three-dimensional scalar fields

$$p_{3,1,0}(x, y, z) = \frac{8}{(2\pi)^2} \rho^2 \left(1 - \frac{\rho}{6}\right)^2 e^{-2\rho/3} \cos^2 \theta$$

$$p_{3,1,1}(x, y, z) = \frac{4}{(2\pi)^2} \rho^2 \left(1 - \frac{\rho}{6}\right)^2 e^{-2\rho/3} (1 - \cos^2 \theta)$$

$$p_{3,2,1}(x, y, z) = \frac{3}{(2\pi)^2} \rho^4 e^{-2\rho/3} \cos^2 \theta (1 - \cos^2 \theta)$$

$$p_{3,2,2}(x, y, z) = \frac{3}{(4\pi)^2} \rho^4 e^{-2\rho/3} (1 - \cos^2 \theta)^2$$

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 $\cos \theta = z/\rho$. **Hint:** To avoid divisions by zero, recast the expressions in terms of $(x, y, z)$ explicitly before evaluating any of them numerically.
3.32. 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

\[
p(x, y, z, t) = A \sin \frac{l\pi x}{a} \sin \frac{m\pi y}{a} \sin \frac{n\pi z}{a} \cos \omega_{lmn} t
\]

where \( l, m, \) and \( n \) are positive integers. Obtain several presentations of the pressure distribution inside this box at \( t = 0 \) and at \( t = \pi/\omega_{lmn} \) for several different values of \( l, m, \) and \( n \).

3.20.6 Graphing Vector Fields

3.33. 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

\[
x_{\text{old}} \leftarrow x_0 \quad \text{Set starting point.}
\]
\[
y_{\text{old}} \leftarrow y_0 \quad \text{Choose step.}
\]
\[
ds \leftarrow \text{chosen step size} \quad \text{Start loop to generate steps.}
\]
\[
\text{LOOP}
\]
\[
\begin{align*}
F_X & \leftarrow F_x(x_{\text{old}}, y_{\text{old}}) \quad \text{Calculate components of force.} \\
F_Y & \leftarrow F_y(x_{\text{old}}, y_{\text{old}}) \\
F_M & \leftarrow \sqrt{F_X^2 + F_Y^2} \quad \text{Calculate magnitude of force.} \\
x_{\text{new}} & \leftarrow x_{\text{old}} + ds*F_X/F_M \quad \text{Step \( dx \) along force line.} \\
y_{\text{new}} & \leftarrow y_{\text{old}} + ds*F_Y/F_M \quad \text{Draw segment of force line.} \\
\end{align*}
\]
\[
\text{Draw line from (x_{\text{old}}, y_{\text{old}}) to (x_{\text{new}}, y_{\text{new}})} \quad \text{Shift focus to new point.}
\]
\[
x_{\text{old}} \leftarrow x_{\text{new}} \quad \text{Repeat.}
\]
\[
y_{\text{old}} \leftarrow y_{\text{new}} \quad \text{EXIT_LOOP WHEN DONE}
\]
\[
\text{END_LOOP}
\]

Cast the loop in this algorithm as a MATLAB procedure \texttt{trace_field} that would be called with a statement like

\[
\text{trace_field}(x_{\text{old}}, y_{\text{old}}, ds, n)
\]

where \( x_{\text{old}} \) and \( y_{\text{old}} \) convey the desired starting point for a field line, \( ds \) conveys the desired step size, and \( n \) 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

\[
\begin{align*}
F_x(x, y) &= \frac{x - 1.5}{[(x - 1.5)^2 + y^2]^{3/2}} - \frac{x + 1.5}{[(x + 1.5)^2 + y^2]^{3/2}} \\
F_y(x, y) &= \frac{y}{[(x - 1.5)^2 + y^2]^{3/2}} - \frac{y}{[(x + 1.5)^2 + y^2]^{3/2}}
\end{align*}
\]

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. \textit{Hint:} 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. \textit{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 $y_{\text{new}}$ deter-
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 $(x_{\text{pred}}, y_{\text{pred}})$.

3.34. Since equipotential curves are perpendicular to the field lines representing the associated (conser-
vative) 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 a MATLAB procedure `trace_equipot` that would be called with a statement like

\[
\text{trace_equipot( x_{\text{old}}, y_{\text{old}}, ds, n )}
\]

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.

3.35. Sometimes, interest can be focussed on the magnitude of the vector field rather than on the full
field. A command like

\[
\text{MAG = sqrt( Hx.^2 + Hy.^2 )}
\]

for example, will generate an array MAG containing the magnitudes of the magnetic field at each
available grid point. This array then conveys a scalar field that can be displayed by any of the
methods described in Section 3.13. Explore the magnetic field given in Eq. (3.16) in this way.

3.36. Read the MATLAB manuals about the procedure `quiver3`. Then use `quiver3` to explore the
velocity field

\[
v(x, y, z) = \omega( -y \hat{i} + x \hat{j} ) + \alpha \hat{k}
\]

(which happens not to depend on $z$) for various values of $\omega$ and $\alpha$. Then write a paragraph or
two describing the field. 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. Hint: In
establishing a scale, you may find it useful to assume a length scale $a$, recast the field in the form

\[
v = a\omega \left( -\frac{y}{a} \hat{i} + \frac{x}{a} \hat{j} \right) + \alpha \hat{k} \implies \frac{v}{a\omega} = \left( -\frac{y}{a} \hat{i} + \frac{x}{a} \hat{j} \right) + \frac{\alpha}{a\omega} \hat{k}
\]

and try to draw graphs of the field $v/a\omega$ as functions of the parameter $\alpha/a\omega$ in the space whose
axes are labeled $x/a$, $y/a$, and $z/a$. 
### 3.20.7 Animation

**3.37.** 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

\[ y(x, t) = \sum_{n=1}^{\infty} A_n \sin \frac{n\pi x}{l} \cos \frac{2\pi nt}{T} \]

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

\[ y(x, 0) = \sum_{n=1}^{\infty} A_n \sin \frac{n\pi x}{l} \]

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 $1/n$.

**3.38.** 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

\[ f(x, t) = e^{-\left(x - vt\right)^2} \quad \text{and} \quad f(x, t) = \sin(x) \cos(t) \]

but feel free to invent others of your own choosing. Here, we suppose that $x$ and $t$ have been cast in appropriate dimensionless units.

**3.39.** 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 3.16.

### 3.20.8 Supplementary Exercises

**3.40.** Write a program to (a) accept as input an arbitrary 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 6

Introduction to MAXIMA

Note: All program (.mac) files referred to in this chapter are available in the directory $HEAD/maxima, 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 MAXIMA’s default search path. If so, the files can be found by MAXIMA 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.

MAXIMA, a large program for manipulating expressions symbolically, is a descendant of one of the earliest computer algebra systems, specifically of MACSYMA, which originated in the late 1960s at the Massachusetts Institute of Technology.\(^1\) MAXIMA can perform symbolic algebra; evaluate derivatives, integrals, and Taylor series; solve algebraic and differential equations; manipulate complex numbers; find eigenvalues and eigenvectors; produce high-resolution two- and three-dimensional graphs; and accomplish numerous other tasks. In this chapter, we describe some ways by which expressions can be defined for MAXIMA and then illustrate some of the statements by which MAXIMA can be instructed to manipulate these expressions. Users should also work through the on-line MAXIMA primer (Section 6.2) and examine several of the built-in demonstrations and examples (also Section 6.2). Further details can be found in printed MAXIMA documentation (Section 6.18) and in on-line help messages accessible from within MAXIMA itself (Section 6.2). We shall refer to this documentation collectively as the MAXIMA manuals.

A typical installation of MAXIMA will make available three different user interfaces at different levels of sophistication. For the most part, we ignore the interface that runs directly in a shell window to the operating system (OS-shell or the maxima window) and limit ourselves to the two interfaces that exploit pop-up windows directly controlling the underlying program: a command-line interface (CLI, most often referred to as the xmaxima window), which—similar to the OS shell—provides coarse resolution and displays equations as they would be constructed with a typewriter, and a graphical user interface (GUI, most often referred to as the wxmaxima window), which displays equations in a more elegant form, admits display of Greek letters and other symbols, and offers menu-based assistance for those who have trouble remembering the available commands. In this chapter, we confine ourselves to describing the structure of statements without addressing the ways in which the GUI facilitates assembling them. While the statements we describe can be entered in any interface, for the most part we work in the GUI.

\(^1\)MACSYMA, which is no longer available, was a proprietary—and expensive—program but MAXIMA is freely available for many operating systems, including Unix, Linux, Windows, Macintosh, and even Android. (See Appendix Z for full contact information.) The terms of the GNU General Public License (GPL) to which users of MAXIMA are subject are explained in the Local Guide.
6.1 Beginning a MAXIMA Session

Detailed instructions for initiating a session with MAXIMA will be found in the Local Guide. In general, MAXIMA will be started either (1) by typing an appropriate command (e.g., `maxima` for the in-shell interface, `xmaxima` for the CLI, `wxmaxima` for the GUI) at the prompt from the operating system,\(^2\) (2) by double-clicking the left mouse button on an appropriate icon\(^3\) on the desktop, or (3) by selecting X-MAXIMA (for the CLI) or WX-MAXIMA (for the GUI) from a menu. Presently, a window labeled Xmaxima console (for the CLI) or wxMaxima (for the GUI) will appear. This window will offer a variety of menus. In the CLI, the window will display a message about MAXIMA and then the prompt (%i1), which will also label the first input line (i-line); in the GUI, the window will be blank, the label (%i1) for the first input line appearing only after that line has been submitted to MAXIMA for execution, though the string --> will appear as the line is being typed.\(^4\) In either case, MAXIMA is ready to accept a succession of statements to control what MAXIMA does on your behalf. We must from the outset be aware that

1. Internally, MAXIMA is case sensitive.

2. Lines of MAXIMA code entered at its prompt can be terminated in any of three ways, depending on whether we want to

   (a) submit the statement for execution and instruct MAXIMA to display its response, in which case we terminate the statement with a semicolon (;),

   (b) submit the statement for execution and instruct MAXIMA to suppress explicit display of its response, in which case we terminate the statement with a dollar sign ($), or

   (c) continue entering the statement on a new line, in which case we simply type \(<\text{ENTER}\rangle\) on the main keyboard (no semicolon or dollar sign) and continue typing the statement, terminating it with a semicolon or a dollar sign when the full statement has been typed.

Once the full statement has been typed and terminated with a semicolon or a dollar sign, the statement is submitted for execution by typing the character from the proper row and column of Table 6.1. Whether the output is actually displayed or not, the output line (o-line) will be given the label (%o\(n\)), where \(n\) is the appropriate sequential number corresponding to the number on the i-line.\(^5\)

3. To place two or more statements in a single physical line, terminate each statement with a semicolon or a dollar sign but hold off actually submitting the statement for execution until the final statement has been typed and terminated. All statements will be executed and each output line (o-line) will be assigned its own label, whether or not the line is displayed. Further, the label on the next i-line will have a number incremented by one from that of the last o-line.

4. Selecting ‘Interrupt’ from the File menu (CLI) or the MAXIMA menu (GUI) causes MAXIMA to abort its present activity and return to a top-level input line, though there may be a short delay because MAXIMA takes time to clean up its workspace before actually returning to its top level. Alternatively, typing \(<\text{CONTROL}/g\rangle\) (CLI) or \(<\text{CONTROL}/G\rangle\) (GUI) will achieve the same objective.

5. Selecting ‘Exit’ from the File menu, double-clicking ML on the MAXIMA icon in the extreme upper left corner of the MAXIMA window, or clicking ML on the ‘\(\times\)’ button in the upper right corner of the MAXIMA window causes MAXIMA to terminate execution, returning control to the computer’s operating system. You may be asked whether to confirm or cancel this action before the action is actually taken.

\(^2\)In these commands to the operating system, case may be important.

\(^3\)The icon for the CLI is likely to be named X-MAXIMA; the icon for the GUI is likely to be named WX-MAXIMA.

\(^4\)The MAXIMA system variable `inchar` defines the character used to label input lines. By default, `inchar` has the value i, but that default value can be changed by assigning a different string to `inchar`. See Section 6.6.

\(^5\)The MAXIMA system variable `outchar` defines the character used to label output lines. By default, `outchar` has the value o, but that default value can be changed by assigning a different string to `outchar`. See Section 6.6.
6.2. **ON-LINE HELP**

Table 6.1: Characters used to submit statements to MAXIMA. Here, \( \langle \text{ENTER} \rangle \) refers to the \( \langle \text{ENTER} \rangle \) key in the main part of the keyboard, \( \langle \text{NUM/ENTER} \rangle \) refers to the \( \langle \text{ENTER} \rangle \) key in the numeric keypad, and \( \langle \text{CONTROL/ENTER} \rangle \) means to hold down the \( \langle \text{CONTROL} \rangle \) key while typing the \( \langle \text{ENTER} \rangle \) key in the main part of the keyboard.

<table>
<thead>
<tr>
<th>OS-shell</th>
<th>X-MAXIMA</th>
<th>WX-MAXIMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Windows</td>
<td>( \langle \text{ENTER} \rangle ) or ( \langle \text{CONTROL/ENTER} \rangle )</td>
<td>( \langle \text{ENTER} \rangle ) or ( \langle \text{NUM/ENTER} \rangle ) or ( \langle \text{CONTROL/ENTER} \rangle )</td>
</tr>
<tr>
<td>LINUX, UNIX, MAC</td>
<td>( \langle \text{ENTER} \rangle ) or ( \langle \text{NUM/ENTER} \rangle ) or ( \langle \text{CONTROL/ENTER} \rangle )</td>
<td>( \langle \text{ENTER} \rangle ) or ( \langle \text{CONTROL/ENTER} \rangle )</td>
</tr>
</tbody>
</table>

6. In the course of a lengthy MAXIMA session, MAXIMA’s workspace may become cluttered with things no longer needed. Sometimes, we would like to restore MAXIMA to its initial state. The simplest way to achieve this objective is to exit from MAXIMA and then restart it. To restore MAXIMA (more or less) to its state immediately upon launch *without exiting and restarting*, invoke any of the following:

- Execute the pair of statements
  
  \[
  (%i25) \text{reset(); kill(all)}$
  
  on a single line in the listed order. Together, these statements will remove all defined labels and reset most global variables and options to their default values.
- Select ‘Restart Maxima’ from the MAXIMA menu in the *wxMaxima* window or ‘Restart’ from the File menu in the *Xmaxima* window.
- Select *Clear Memory* from the MAXIMA menu in the *wxMaxima* window or ‘kill(all)’ from the MAXIMA→CLEAR MEMORY menu in the *Xmaxima* window.

6.2 **On-Line Help**

MAXIMA makes several sorts of on-line help available:

1. A description of any specific command will be displayed if the command \( ? \text{CommandName} \);, e.g., \( ? \text{expand} \);, is typed at the MAXIMA prompt. Note that the space after the question mark is mandatory.

2. Users can peruse on-line manuals by selecting whatever of ‘MAXIMA Manual’, ‘Xmaxima Manual’, ‘MAXIMA Help’ and/or ‘wxMaxima Help’ is available in the HELP menu. Further, the *MAXIMA Primer* pops up automatically when the CLI is invoked and can, alternatively, be displayed by opening the file \$\text{MAXHEAD}/maxima/5.36.1/xmaxima/intro.html\) in a browser.\(^6\)

\(^6\)Here, \$\text{MAXHEAD}\) identifies the top-level directory in which files associated with MAXIMA are stored; it is defined for your site in the *Local Guide*. Further, the folder 5.36.1 in this path identifies the version of MAXIMA in use and will be different if an alternative version is installed in your system.
3. The HELP menu in the GUI offers the option to display built-in examples of some of the MAXIMA functions. For example, in the GUI, selecting ‘Example...’ from the HELP menu brings up a pop-up box in which you can type the name of a command, e.g., diff, and then click ML on ‘OK’ to display one or more examples of the use of the specified command. Alternatively, in either interface, the statement

\[ \text{example( ExampleName );} \]

will also display the specified example. However invoked, the statement effects a reading of one or more statements from the selected example file and displaying the output of each statement, doing so without user intervention at any point. Some of these examples are quite long. Scrolling through the individual components after they have all been displayed is, of course, possible. The statement \[ \text{example();} \] returns a long list of all available examples.

Your attention is drawn particularly to the examples complex, diff, expand, integrate, matrices, radcan, ratexpand, ratsimp, solve, trig, and ode2.

4. A built-in demonstration having the name DemonstrationName and illustrating the use of a particular MAXIMA capability can be invoked by typing the statement

\[ \text{demo( DemonstrationName );} \]

e.g., \[ \text{demo( cf );} \], on any input line. When this statement is invoked on an available demo, the components are displayed one at a time. To continue to the next component, the user must type ; and then \{ENTER\} on the numeric keypad or \{CONTROL/ENTER\} on the main keyboard each time the display pauses.

In the CLI, a demonstration may be stopped before it comes to its natural end by selecting ‘Input prompt’ (equivalent to typing \{ALT-s\}) or ‘Restart’ from the File menu; in the GUI, a demonstration may be stopped prematurely by selecting ‘Restart Maxima’ from the Maxima menu.

There are very few useful built-in demonstrations. Note, however, that most built-in examples can be displayed line by line rather than non-stop by typing the statement

\[ \text{demo( "ExampleName.mac" );} \]

5. By selecting ‘Tutorials’ in the HELP menu in the GUI, the user can bring up the wxMaxima documentation window, which provides links to numerous tutorials.

6. The statements

\[ \text{describe( CommandName );} \quad \text{or} \quad ? \text{ CommandName} \]

e.g., \[ \text{describe(diff);} \quad \text{or} \quad ? \text{ diff;} \] typed at any input line will bring up either a description of the command or a list of topics available for the specified command and offer the option of viewing one or more of those topics.
6.3 Basic Entities in MAXIMA

At the very beginning, we point out that MAXIMA distinguishes among a variety of different entities, including

- expressions, lists, and arrays, which are the primary quantities with which MAXIMA works;
- functions, which accept zero, one, or more arguments as input and return various outputs, thereby providing the means by which expressions, lists, and arrays are manipulated;
- option variables (some of which are called plot options), whose values—default or user-specified—modify the behavior of associated functions;
- system variables, which are maintained by MAXIMA but contain possibly useful information;
- special symbols, including \%pi for \(\pi\), \%i for \(\sqrt{-1}\), \%e for \(e\), \%gamma for the Euler-Mascheroni constant, \text{inf} for \(+\infty\), and \text{minf} for \(-\infty\);
- keywords, which may appear among the arguments to functions and modify the output produced by the functions.

6.4 Variable Names in MAXIMA

Legal variable names, sometimes called identifiers, begin with an alphabetic character (upper or lower case), a percent sign, an underscore character, or a backslash followed by a numeral. Later characters may be any of these characters, but only an initial numeral requires a preceding backslash. Variable names can be of any length. Lower case and upper case letters are treated as distinct, so the variable \text{ABC} is different from the variable \text{abc}.

Greek letters cannot be displayed as such in the CLI (\text{xmaxima} window), though \text{alpha} or \text{Alpha}, for example, can certainly be the name of a variable. In the GUI (\text{wxmaxima} window), with three exceptions,\(^8\) spelling out the name of a Greek letter in lower-case letters in an i-line will result in the display of the lower-case Greek letter in the associated o-line, e.g., \text{alpha} in an i-line will produce \(\alpha\) in the corresponding o-line. Also in the GUI, if the initial letter of the name of a Greek character in an i-line is made upper case, the corresponding upper-case Greek letter is produced in all cases (i.e., with no exceptions).\(^9\)

\text{Note particularly that MAXIMA will by default assume that any variable to which no explicit value has been assigned represents a real quantity.}

6.5 Expressions in MAXIMA

Expressions are written in MAXIMA using a syntax very much like that used in ordinary algebra and are supplied merely by typing the expression at a input line,\(^11\) e.g.,

\(^8\) The three exceptions are (1) \text{gamma} produces an upper-case \(\Gamma\), (2) \text{lambda} is left as \text{lambda}, and (3) \text{psi} produces an upper-case \(\Psi\).

\(^9\) Whether the initial letter is lower or upper case, preceding the Greek name with a percent sign \% will display the corresponding lower or upper case Greek letter in all cases, though \%gamma, displayed \(\gamma\), is a system variable that is bound to the value 0.577215..., the value of the Euler-Mascheroni constant, \%pi, displayed \(\pi\), is a system variable that is bound to the value 3.14159265..., and \%phi, displayed \(\phi\), is a system variable bound to the value of the golden mean \((1 + \sqrt{5})/2 = 1.618033...\).

\(^10\) In the GUI, preceding the (spelled-out) name of a Greek letter with a backslash \textbackslash in an input line will generate the corresponding upper- or lower-case Greek letter in the output line, though with the same three exceptions enumerated above (\text\%gamma produces \(\Gamma\), \text\%lambda produces \text{lambda}, and \text\%psi produces \(\Psi\)).

\(^11\) In this section, we use non-sequential statement numbers so as 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.
\((\%i12) \ x + a \ * \cos(b*x)\);

to which, since the expression was terminated with a semicolon rather than a dollar sign, MAXIMA will respond\(^\text{12}\)

\((\%o12) \ a \cos(bx) + x\)

The symbols +, -, *, /, and ^ are used for addition, subtraction, multiplication, division, and exponentiation.\(^\text{13}\) Except for those in quoted strings, spaces are ignored by MAXIMA and may be freely inserted for legibility. Functions available for use in expressions include

- General functions, e.g., abs(), float(), bfloat(), fix(), round(), sqrt(), signum(), unit_step(), and random().
- Trigonometric, hyperbolic and exponential functions and their inverses, e.g., sin(), cos(), tan(), atan(), atan2(), sinh(), cosh(), tanh(), atanh(), exp(), and log().
- Special functions, e.g., Bessel and modified Bessel functions, Airy functions, elliptic integrals, gamma and beta functions, error function, Legendre polynomials, Hermite polynomials, and Laguerre polynomials.

Full details on these and other functions, including information about the form of their arguments and option variables that affect their behavior, will be found in the MAXIMA manuals.

Internally, MAXIMA sees simple expressions as lists. The first element (known as part zero of the expression) indicates the operator involved, and the remaining elements (known as parts one, two, ...) indicate the operands. Thus, for example, the two expressions

\[ a + b + c + d \]
\[ a \times b \times c \]

are stored internally as the lists

\[[+], [a], [b], [c], [d]\]
\[[\times], [a], [b], [c]\]

respectively. More complicated expressions, such as

\[ a + b - c + d \]
\[ a + b(c + d) + e \]

are stored as lists of lists, being

\[[+], [a], [b], [-], [c], [d]\]
\[[+], [a], [+], [b], [+], [c], [d]\], [e]\]

respectively. As a final example, which anticipates a later need, suppose the expression of interest is

\[(x + 3)^2 (x - a) (x + a)\]

We would generate MAXIMA’s internal representation in several steps finding

\[[\times], [(x+3)^2], [x-a], [x+a]\]
\[[\times], [[-], [x+3], [2]], [+, [x], [-a]], [+, [x], [a]]\]
\[12\text{Output of MAXIMA on different platforms may not be character for character identical to what is displayed in this book. Also note that we will occasionally rearrange the display of MAXIMA output lines to improve legibility. No substance is lost in those rearrangements.}\n\[13\text{Following the pattern in FORTRAN, MAXIMA also recognizes the symbol ** for exponentiation.}\]
Here, the first line shows parts 0, 1, 2, and 3 of the original expression while the second line reveals that part 1 of the original expression is itself a list having part 0 (part 0 of part 1 of the original expression; \(\ast\)), part 1 (part 1 of part 1; \(x+3\)), and part 2 (part 2 of part 1; 2). Indeed, 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 3 while part 1 of part 2 of part 1 of the original expression is \(a\). These relationships can be involved and confusing, but some manipulations depend critically on being able to identify the parts, subparts, subsubparts, and \ldots{} of a complicated expression. Note that, in many cases, part 0 of an expression or subexpression is the operator to be applied to the remaining parts of the expression.

6.6 Binding Expressions 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. MAXIMA provides for this shorthand in two ways. First, as a session proceeds, MAXIMA maintains an internal record of the contents of each i-line and each o-line and identifies each such item with the label on the line on which it appeared. Each item can then be referred to in subsequent statements by that label (\(%i?\) or \(%o?\), where \(?\) stands for the number assigned to the line. For example, a later request that MAXIMA square the expression on line \(%i12\) in the previous section could be communicated to MAXIMA with the statement

\[
(%i18) \ %i12^2; \\
(%o18) \ (a \cos(bx) + x)^2
\]

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 chosen variable. The statement,\(^{14}\)

\[
(%i20) \ \textbf{bigexp} : x + a \ast \cos(b \ast x); \\
(\textbf{bigexp}) \ \ a \cos(bx) + x
\]

for example, binds the expression following the colon to the variable \textbf{bigexp}. Subsequently, whenever that variable is included in an expression, it will automatically be replaced by the expression to which it has been bound, and the example on line \(%i18\) above might then have been written

\[
(%i30) \ \textbf{bigexp}^2; \\
(%o30) \ (a \cos(bx) + x)^2
\]

Binding an expression to a variable is preferred over using line labels because the chosen variable can have mnemonic significance, and it is decidedly easier to remember the variable to which a particular expression has been bound than to remember the label on an i-line or o-line, especially after the associated line has scrolled out of the \textit{MAXIMA} window.

Even if a mnemonic name is used, we can remind ourselves of the value bound to a particular variable simply by asserting its name. The statement

\[
(%i47) \ \textbf{bigexp};
\]

for example, will request MAXIMA to display the current value of \textbf{bigexp} on \(%o47\). Further, if we have no further need for a particular variable and wish to tell MAXIMA to forget that it once had a value bound to the variable, we could enter the statement

\(^{14}\text{In recent versions of MAXIMA, the label o-label that would normally be displayed when a variable is explicitly assigned a value may be replaced with the name of that variable, though the corresponding o-label—here \(%o20\)—will still be assigned and may be used to refer to that output in subsequent statements. Throughout \textit{CPSUP}, we will probably not consistently use either format of the label on an output line.}
\texttt{(%i83) remvalue( bigexp );}

...to remove the variable and free some of MAXIMA's workspace.

When special functions of significance to a particular session would be useful, we might exploit the MAXIMA operator \texttt{:=} to append any desired functions to the built-in functions. For example, the statement

\texttt{(%i121) spectrig(x) := 1 + sin(x)^2;}

typed at any input line defines the function \texttt{spectrig()} for subsequent use just as we would use any of MAXIMA's built-in functions. Further, if we have no further need for a particular function and wish to tell MAXIMA to forget that it once associated a function with a particular variable, we could enter the statement

\texttt{(%i175) remfunction( spectrig );}

...to remove the function and free some of MAXIMA's workspace.

\section*{6.7 Using MAXIMA ...}

At base, a session with MAXIMA involves entering one or more expressions and then instructing MAXIMA to process these expressions in some way. The following transcripts of MAXIMA sessions illustrate some of the simpler manipulations. The transcripts are presented in two columns, the one on the left recording the statements submitted to MAXIMA and MAXIMA's responses, the one on the right containing explanatory comments.

As a statement is being entered at a MAXIMA prompt, both in the CLI and in the GUI, the arrow, home, and end keys can be used to move the cursor around in the line and on the screen, the backspace and delete keys can be used to remove characters, and all printing keys can be used to insert characters at the point of the cursor. Further, the cursor need not be at the end of a line for the line to be submitted for execution. After a statement has been submitted for execution, the cursor can be moved back into that line, edits can be made, and the line can be resubmitted. Depending on the particular implementation of MAXIMA in use, the record of the activity on the screen may become unclear in one way or another. For example, in the CLI in Windows, the original i-line label will remain on the screen and continue to identify the \textit{original} content of the line even though the statement displayed on the line will reflect the edits made. In contrast, in the GUI in Windows, the original i- and o-line labels will disappear from the screen altogether and be replaced by new labels, even though the no-longer-displayed labels remain identified with their \textit{original} meaning. Editing and resubmitting existing lines may create a confusing, out-of-order on-screen record of the unfolding activity.

To keep the several segments in this section somewhat separated from one another, we shall periodically either (1) exit from MAXIMA as described in item 5 of Section 6.1 and reenter MAXIMA or (2) restart MAXIMA as described in item 6 of Section 6.1, thereby clearing MAXIMA's workspace and restarting the i- and o-line labels. This action is taken primarily to avoid having a single string of line labels throughout this section. The action is 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 \textit{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 MAXIMA.

Finally, note that, throughout this section, various components of MAXIMA (variable names, functions, option variables, plot options, system variables, operators, ...) are introduced simply by presenting statements in which they are used, frequently with \\textit{brief} comment. Details on all mentioned entities and on many others can be found in the MAXIMA manuals.
6.7. USING MAXIMA ...

6.7.1 ... for Arithmetic

We illustrate first some of the ways in which MAXIMA can be instructed to perform simple arithmetic operations. Consider, for example, the statements

(%i1) 2 + 3;
(%o1) 5

Enter a simple numeric expression. MAXIMA returns a value, *doing* the arithmetic.

(%i2) 4/6;
(%o2) \frac{2}{3}

Enter another expression. MAXIMA does not automatically generate a floating point evaluation but *did* reduce the fraction.

(%i3) sqrt(9);
(%o3) 3

Compute \( \sqrt{9} \).

(%i4) sqrt(12);
(%o4) 2\sqrt{3}

Compute \( \sqrt{12} \). MAXIMA simplifies the expression but keeps it exact.

(%i5) factorial(9);
(%o5) 362880

Find \( 9! = 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 \), with *factorial*.

(%i6) 9!;
(%o6) 362880

Find \( 9! \) with postfix operator \(!\).

(%i7) 9!!;
(%o7) 945

Find \( 9 \) double-factorial, \( 9!! = 9 \times 7 \times 5 \times 3 \times 1 \).

(%i8) float( %o2 );
(%o8) 0.6666666666666666

Ask for floating value. By default, the value is computed as a double-precision value. 

When the *option variable* \( \text{fpprintprec} \) has a value between 2 and 16 inclusive, that value determines the number of digits actually printed by \( \text{float} \). For other values of this variable (including the default value 0), 16 digits will be printed. Note, however, that trailing zeroes will be suppressed.

(%i9) bfloat( %o2 );
(%o9) 6.66666666666666666666667b-1

Ask for bigfloat precision floating value. That MAXIMA has returned such a value is conveyed by the \( b \) preceding the exponent in the output. The number of digits actually displayed is controlled by \( \text{fpprintprec} \) unless \( \text{fpprintprec} \) exceeds the option variable \( \text{fpprec} \), in which case \( \text{fpprec} \), with default value 16, stipulates the number of digits to be displayed. Again trailing zeroes are suppressed.

(%i10) [fpprec, fpprintprec];
(%o10) [16, 0]

Display current values of controlling option variables.

(%i11) fpprec:24$ fpprintprec:30$

Reset those values. Here, the dollar signs in line (%i11) suppress display but not behind-the-scenes evaluation of MAXIMA’s response. Note that the labels \%o11 for \( \text{fpprec} \) and \%o12 for \( \text{fpprintprec} \) are still assigned.

(%i12) bfloat(t);
(%o12) 6.66666666666666666666667b-1

Display \( \text{bfloat} \) again, this time with the number of digits controlled by \( \text{fpprec} \), since \( \text{fpprintprec} > \text{fpprec} \).

(%i13) reset( fpprec, fpprintprec )$

Reset default values of option variables.
Ask for floating value of $\sqrt{12}$.

We have, of course, illustrated only some ($\ast, /, \sqrt{}, \text{factorial}, !, !!, \text{float}, \text{bfloat}$) of the operators and functions that might be used to persuade MAXIMA to manipulate with numbers and only two ($\text{fpprec}$ and $\text{fpprintprec}$) of the option variables whose values affect the behavior of these functions. You should make it a point to read about these and other functions, option variables, and operators in the MAXIMA manuals.

6.7.2 ... for Algebra

MAXIMA 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 in the o-line that the order of factors may—and here does—differ from that entered. MAXIMA arranges expressions internally to facilitate their subsequent manipulation.

Expand the expression. Here, the symbol $\%$ refers to the most immediately past o-line.

Factor the expression on line $($%o2$)$.

We note, of course, that this factoring hasn’t quite returned us to our starting point. To achieve that end, we need to multiply out the second and third factors without expanding the first. In essence, we will have to carve the expression apart, manipulate with its parts, and then reconstruct the expression. As discussed in Section 6.5, MAXIMA sees this expression as lists of lists. Indeed, we can persuade MAXIMA to reveal that structure by exploiting the MAXIMA command $\text{part}$ in the statement

or we could exploit not only the function $\text{part}$ but also the command $\text{print}$ and MAXIMA’s ability to construct a loop in the statement$^{15}$

Using these capabilities, we might achieve the entire objective with the single nested statement

$^{15}$In more recent versions of MAXIMA, a colon : can be substituted for from in this statement. We will adhere to the older syntax, which still works in these more recent versions.
in which we extract parts 1, 2, and 3 from \( %o3 \), expand the product of parts 2 and 3, and then reconstruct the expression by multiplying that result by part 1. This time, we have succeeded in reconstructing the original expression. As here illustrated, beating MAXIMA's version of a complicated expression into a more conventional form may require a fair bit of creative massaging.

We continue with some additional manipulations. First, we set the constant \( a \) in the previous result to the specific value 2 with the statement

\[
(%i17) \text{ev( } \%, \text{ a=2 );}
\]

\[
(%o17)
\]

Evaluate \( %o6 \) with \( a = 2 \). The simpler statement \( \%, \text{ a=2; } \) would accomplish the same end without invoking \text{ev}.

Then we find a partial fraction expansion of the reciprocal of this polynomial with the statement

\[
(%i18) \text{partfrac( 1/%, x );}
\]

\[
(%o18)
\]

Find partial fraction expansion of \( 1/%o7 \), expanding with respect to \( x \).

Numerous rewrites of this expression are now possible. For example, replacing \( x \) with \( 3z \) would be accomplished with the command \text{subst}, as for example in the statement

\[
(%i19) \text{subst(3*z, x, % );}
\]

\[
(%o19)
\]

Substitute \( 3z \) for \( x \) in \( %o8 \).

To illustrate a more complicated recasting, suppose we wanted to square the denominator of the second term in \( %o8 \) without affecting any other term. The statement \text{expand}(%o8) does more than we want; it will expand \textit{every} denominator. (Try it.) We need a more sophisticated operation that rearranges only the second term (part 2 of the original expression). We extract that term, carve it up, manipulate that term alone, and substitute the result of that manipulation into the original expression as the second term. We identify the proper part with the statements

\[
(%i10) \text{part( %o8, 2 );}
\]

\[
(%o10)
\]

Extract second part of \( %o8 \).

\[
(%i11) \text{part( %, 2 );}
\]

\[
(%o11)
\]

Extract second part of second part of \( %o8 \).

\[
(%i12) \text{part( %, 2 );}
\]

\[
(%o12)
\]

Extract second part of second part of second part of \( %o8 \).

\[
(%i13) \text{part( %o8, 2, 2, 2 );}
\]

\[
(%o13)
\]

Alternatively, extract the second part (third 2) of the second part (second 2) of the second part (first 2) of \( %o8 \) in one step.

Then we expand the extracted piece of the denominator of the second term in the original expression and substitute it into \( %o8 \) as part 2.2.2 with the statement

\[
(%i14) \text{substpart( expand( % ), %o8, 2, 2, 2 );}
\]

\[
(%o14)
\]

See item 6 in Section 6.1.

Note once again that MAXIMA has a mind of its own regarding the order of terms in an expression.

We have, of course, only mentioned a few (expand, \text{factor}, part, \text{ev}, \text{partfrac}, subst, substpart) of the numerous functions MAXIMA makes available for manipulating expressions algebraically. We have also mentioned the function print and the control structure for-do. The
6.7.3 ... for Complex Variables

MAXIMA is also capable of manipulating complex variables. We illustrate some of MAXIMA's main capabilities in this category with the statements

\begin{verbatim}
(%i1) fct : sin(x+y*%i);
(%o1) sin(%iy + x)

(%i2) realpart( fct );
(%o2) sin(x) cosh(y)

(%i3) imagpart( fct );
(%o3) cos(x) sinh(y)

(%i4) conjugate( fct );
(%o4) -sin(%iy - x)

(%i5) (x+%i*y)/(x-%i*y);
(%o5) (x + %i*y) / (x - %i*y)

(%i6) [realpart(%), imagpart(%)];
(%o6) [x^2 - y^2, y^2 + x^2]

(%i7) abs( %o5 );
(%o7) |%i y + x| / |%i y - x|

(%i8) cabs( %o5 );
(%o8) 1

(%i9) q : realpart( %o5 ) + %i*imagpart( %o5 );
(%o9) x^2 - y^2 / y^2 + x^2 + 2 %i xy / y^2 + x^2

(%i10) q1 : carg( q );
(%o10) atan2(2%xy, x^2 - y^2 / y^2 + x^2)
\end{verbatim}

Find the absolute value of \( \%o5 \). MAXIMA assumes \( x \) and \( y \) are real but notes \( \%i = \sqrt{-1} \), so punts on the evaluation.

Find complex absolute value of \( \%o5 \). \texttt{cabs} also assumes \( x \) and \( y \) to be real but knows how to deal with \( \%i = \sqrt{-1} \). \texttt{cabs} also simplifies the result obtained from simply the square root of the sum of the squares of the real and imaginary parts!

Recast \( \%o5 \) as sum of real and imaginary parts.

Find complex argument of \( q \). Note that \texttt{carg} gives a different—but presumably equivalent—result if used in the form \texttt{carg(\%o5);}.

In the evaluation of \texttt{atan2}, only the ratio of the two arguments is important. Recognizing that the first and second arguments of the result in \( q1 \) are parts 1 and 2 of \( q1 \), we can simplify the result just obtained by invoking the statement...
6.7. USING MAXIMA

(%i11) [p1:part(q1,1), p2:part(q1,2)];
Extract and name arguments of q1.

(%o11) \begin{bmatrix}
2xy & x^2 - y^2 \\
y^2 + x^2 & y^2 + x^2 
\end{bmatrix}

(%i12) p1/p2;
Calculate ratio of those arguments.

(%o12) \frac{2xy}{x^2 - y^2}

(%i13) q2: atan2(num(%),denom(%));
Reconstruct arctangent using numerator and denominator of the ratio of the two original arguments.

(%o13) \text{atan2}(2xy, x^2 - y^2)

(%i14) x + %i*y;
Define a complex number.

(%o14) %iy + x

(%i15) polarform( %o14 );
Convert to polar form.

(%o15) \sqrt{y^2 + x^2} \%e^{\%i \text{atan2}(y,x)}

(%i16) r*exp(%i*theta);
Define another complex number.

(%o16) r \%e^{\theta}

(%i17) rectform(%o16 );
Convert to rectangular form.

(%o17) %ir \sin(\theta) + r \cos(\theta)

(%i18) reset()$ kill(all)$
See item 6 in Section 6.1.

We have, of course, only mentioned some (\texttt{realpart}, \texttt{imagpart}, \texttt{conjugate}, \texttt{abs}, \texttt{cabs}, \texttt{carg}, \texttt{polarform}, \texttt{rectform}) of the functions MAXIMA makes available for manipulating complex numbers. You should make it a point to read about these functions and any associated option variables in the MAXIMA manuals.

6.7.4 ... for Trigonometry

MAXIMA 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

(%i11) trigexp : sin(x)*cos(x); Define a trig expression.
(%o11) \cos(x)\sin(x)

(%i12) exponentialize(trigexp);
Convert to exponential form.

(%o12) \frac{-\%i(\%e^{\%ix} - \%e^{-\%ix})(\%e^{\%ix} + \%e^{-\%ix})}{4}

(%i13) demoivre(%);
Convert to standard trig functions.
(%o13) \cos(x)\sin(x)

(%i14) trigreduce(trigexp);
Recast with multiple angle identities.

(%o14) \frac{\sin(2x)}{2}

(%i15) trigexpand(%);
Expand functions of multiple angles.
(%o15) \cos(x)\sin(x)
Define another trig function.

Expand using trig addition formulae.

Do it again.

Define yet another trig function.

Expand it.

Simplify it. Clearly, trigsimp knows the standard trigonometric identities.

To illustrate the use of a loop once again, let us create a list of the values of the sine of integer multiples of π/2. We use the statements

```maxima
(%i12) for n from 0 thru 10 do a[n]: sin(n*%pi/2);
(%o12) done
(%i13) b : makelist( a[n], n, 0, 10 );
(%o13) [ 0, 1, 0, -1, 0, 1, 0, -1, 0, 1, 0 ]
```

MAXIMA’s symbol %pi stands for π = 3.14159...

We have, of course, only mentioned a few (exponentialize, demoivre, trigreduce, trigexpand, expand, and trigsimp) of the functions MAXIMA makes available for doing trigonometry. We have also illustrated again the use of the control structure for/from-thru/do. Additional important functions include exp, log, logarc, and logcontract. Further, the action of one or more of these functions may be influenced by one or more of the option variables %emode, exponentialize, numer, trigexpandplus, trigexpandtimes, triginverses, logexpand, logarc, demoivre, and halfangles, among others. You should make it a point to read about these commands and option variables in the MAXIMA manuals.

6.7.5 ... for Algebraic Equations

We turn next to illustrating a few MAXIMA commands for manipulating expressions algebraically. To define a polynomial and find its roots, for example, we would use statements like

```maxima
(%i11) poly : expand( (x^2-a^2) * (x+3)^2 );
(%o11) x^4 + 6x^3 - a^2x^2 + 9x^2 - 6a^2 x - 9a^2
(%i12) solve( poly = 0, x );
(%o12) [x = -a, x = a, x = -3]
```

Solve the equation poly = 0 for x. There are three distinct roots. We could alternatively have submitted the statement `solve( poly, x );`, in which case MAXIMA would have understood the equation to be poly = 0.

Show the system variable multiplicities to learn that the first two roots are simple roots and the third is a double root.

Verify first root. Alternatively, we could have submitted the statement poly, x=-a; with the same end result.
6.7. USING MAXIMA ...

(%i5) for rt in %o2 do print( ev(poly,rt) );
Verify all roots, using the for-do structure
to code a loop.
0
0
0
(%o5) done
(%i6) a*x^2 + b*x + c;
Supply a quadratic equation.
ax^2 + bx + c
(%o6) done
(%i7) solve( %, x );
Solve %o6 = 0 for x. Clearly, MAXIMA
knows the quadratic formula.
\[
\begin{align*}
  x &= -\frac{\sqrt{b^2 - 4ac} + b}{2a}, \\
  x &= \frac{-\sqrt{b^2 - 4ac} - b}{2a}
\end{align*}
\]
(%o7) done
(%i8) %, a=3, b=4, c=2;
Evaluate %o7 with indicated values of a, b,
ad c. Note that MAXIMA does not detect
the common factor of 2 in numerator and
denominator of these expressions, though
factor applied to %o8 will effect that sim-
plification.
\[
\begin{align*}
  x &= -\frac{23/2 \cdot i + 4}{6}, \\
  x &= \frac{23/2 \cdot i - 4}{6}
\end{align*}
\]
(%o8) done
We can, of course, always ask for floating point evaluations with the command float, finding here that
(%i9) float(%);
\[
\begin{align*}
  x &= -0.1666666666666667 (2.828427124746191 \cdot i + 4.0), \\
  x &= 0.1666666666666667 (2.828427124746191 \cdot i - 4.0)
\end{align*}
\]
(%i10) expand(%);
\[
\begin{align*}
  x &= -0.4714045207910318 \cdot i - 0.6666666666666666, \\
  x &= 0.4714045207910318 \cdot i - 0.6666666666666666
\end{align*}
\]
In the second statement, we have invoked expand to multiply the initial factor in each solution into
each term of the following binomial.
Finally, let us illustrate MAXIMA’s capability to solve simultaneous linear equations. To define
and solve a representative system, we invoke the statements
(%i11) eqn1 : x - y = 4;
Give MAXIMA a pair of simultaneous lin-
erar equations, binding each to a variable to
facilitate subsequent reference.
\[
\begin{align*}
  x - y &= 4
\end{align*}
\]
(%o11) done
(%i12) eqn2 : x + y = 0;
\[
\begin{align*}
  x + y &= 0
\end{align*}
\]
(%o12) done
(%i13) solve( [eqn1,eqn2], [x,y] );
Solve the system for x and y.
\[
\begin{align*}
  [x = 2, y = -2]
\end{align*}
\]
Note that each solution—here there is only one—is presented as a list. That there are two opening
and closing square brackets anticipates situations in which there are two or more solutions. The
structure of the output would then be [[, [, [, ...]], and the pair of square brackets at the
extreme ends would seem less redundant.
(%i14) eqn1 : a*x+b*y = e$ Try it with a more symbolic system.
(%i15) eqn2 : c*x+d*y = f$
(%i16) solve( [eqn1,eqn2], [x,y] );
\[
\begin{align*}
  x &= \frac{-de - bf}{bc - ad}, \\
  y &= \frac{ce - af}{bc - ad}
\end{align*}
\]
(%o16) done
(%i17) remvalue( eqn1, eqn2 )$ Remove values from eqn1, eqn2.
We have, of course, only mentioned a few (solve, multiplicities) of the functions and system variables MAXIMA makes available for solving algebraic equations exactly. Behind the scenes, solve actually invokes either linsolve or algsys, as appropriate to the equations presented to it. Further, the functions realroots, allroots, and newton can find roots numerically. These functions may be influenced by one or more of the option variables algepsilon, algexact, rootsepsilon, breakup, programmode, realonly, globalsolve, and solveexplicit, among others. You should make it a point to read about these functions, option variables, and system variables in the MAXIMA manuals. Root finding is discussed in greater depth in Chapter 14.

6.7.6 ... for Basic Calculus

Differentiation, indefinite and definite integration, and determination of series representations of symbolic expressions are also among MAXIMA’s capabilities. To evaluate derivatives and integrals, for example, we would use statements like

(%i11) x^2*exp(-x^2);
(%o11) x^2 \times e^{-x^2}

(%i12) diff( %, x );
(%o12) 2x \times e^{-x^2} - 2x^3 \times e^{-x^2}

(%i13) integrate( %, x );
(%o13) -(x^2 - 1) \times e^{-x^2} - x \times e^{-x^2}

(%i14) ratsimp( % );
(%o14) x^2 \times e^{-x^2}

(%i15) factor( diff( %o1, x, 2 ) );
(%o15) 2(2x^4 - 5x^2 + 1) \times e^{-x^2}

(%i16) x/(a^2+x^2)^{3/2};
(%o16) \frac{x}{(x^2 + a^2)^{3/2}}

(%i17) integrate( %, x, 0, inf );
Is a zero or nonzero? n;
(%o17) - \frac{1}{a}

(%i18) integrate( %o6, x, 0, inf );
Is a zero or nonzero? z;
(%o18) \frac{1}{a}

This example has been included to draw attention to the need to be careful in using symbolic manipulating programs, since

\[ \int_0^\infty \frac{x}{(x^2 + a^2)^{3/2}} \, dx = \begin{cases} 
1/|a| & a \neq 0 \\
\text{divergent} & a = 0 
\end{cases} \]  

(6.1)

MAXIMA’s value for this integral is incorrect. The question asked by MAXIMA is correct, but MAXIMA should return 1/|a| for a \neq 0 and something like ‘divergent’ for a = 0.\textsuperscript{16}

\textsuperscript{16}The developers of MAXIMA have been alerted to this bug and will surely have corrected it in future versions of MAXIMA.
Finally (for this section), we point out that MAXIMA can generate series expansions of prescribed functions. To illustrate, we ask MAXIMA to process the statements

\[(%i9) \frac{1}{1+\exp(-x)};\]
\[\text{Enter yet another expression.}\]
\[\frac{1}{%e^{-x} + 1}\]
\[\text{(%o9) }\frac{1}{%e^{-x} + 1}\]

\[(%i10) \text{taylor( } %, x, 0, 5 );\]
\[\text{Expand the expression in a Taylor series in } x \text{ about } x = 0, \text{ retaining terms to 5-th order in } x. \text{ The notation } /T/ \text{ at the beginning of the o-line indicates that the line contains a Taylor series.}\]
\[\text{(%o10) } /T/ \frac{1}{480} x^5 + \frac{-1}{48} x^3 + \frac{1}{4} x + \frac{1}{2} + \ldots\]

\[(%i11) \text{reset()}$ kill(all)$\]
\[\text{See item 6 in Section 6.1.}\]

We have, of course, only mentioned a few (\texttt{diff}, \texttt{integrate}, \texttt{taylor}) of the numerous functions MAXIMA makes available for doing calculus. Important additional functions include \texttt{changevar}, \texttt{gradeff}, \texttt{lddefint}, \texttt{limit}, and \texttt{romberg} (for numerical integration). Further, the action of one or more of these functions may be influenced by one or more of the option variables \texttt{derivabbrev}, \texttt{intanalysis}, \texttt{logabs}, \texttt{rombergtol}, and \texttt{rombergabs}, among others, and MAXIMA provides some information about what it has done in the system variable \texttt{dependencies}. You should make it a point to read about these functions, option variables, and system variables in the MAXIMA manuals.

### 6.7.7 ... for Laplace Transforms

In Section 1.5.2, we defined the Laplace transform \(\tilde{f}(s)\) of a function \(f(t)\) by the integral

\[\tilde{f}(s) = \int_0^\infty e^{-st} f(t) \, dt\]  

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 \textit{differential} equation into a linear \textit{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 \textit{differential} equation is to take its Laplace transform, solve the resulting \textit{algebraic} equation for that transform, and then invert that transform to return from the \(s\)-space of the transform to the \(t\)-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.

Within MAXIMA, the function \texttt{laplace} will calculate a Laplace transform and the function \texttt{ilt} will invert a transform to return to the original space. We illustrate the use of those functions with the statements

\[(%i11) \text{laplace( } \sin(\omega \text{t}), \text{ } t, \text{ } s ) ;\]
\[\text{Calculate the Laplace transform of the function } \sin(\omega t), \text{ where } t \text{ is the primary variable and } s \text{ is the variable in terms of which the transform is to be expressed.}\]
\[\text{(%o1) } \frac{\omega}{s^2 + \omega^2}\]
CHAPTER 6. INTRODUCTION TO MAXIMA

Then, invert the transform to return to the original function. The intermediate question recognizes a distinction based on the value of the parameter $\omega$.

Seek result for the other option in the question asked. Here is another apparent error in MAXIMA: If $\omega$ is a (hard) zero, then the Laplace transform involved ($\%o1$ with $\omega = 0$) is zero and the inverse transform should also be zero.

Calculate the Laplace transform of the function $e^{-at}$.

Calculate the inverse Laplace transform of the function $(s + a)^{-1}(s + b)^{-1}$.

See item 6 in Section 6.1.

6.7.8 ... for Ordinary Differential Equations

Among MAXIMA’s strongest suits is its ability to solve ordinary differential equations (ODEs). We will postpone introducing the more sophisticated of those capabilities until Chapter 11, limiting ourselves now to a quick illustration of the simplest of them. We first must tell MAXIMA that the function we seek—say $x(t)$—depends on the independent variable $t$ with the statement

Then, we tell MAXIMA the differential equation. For the sake of a specific example, we here choose 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 MAXIMA with the statement

Here $\text{diff}(x,t,2)$ specifies the second derivative of $x$ with respect to $t$ while $\text{diff}(x,t)$ specifies the first derivative. The opening tic (’) suppresses immediate evaluation of the derivative (and is recommended even though, in this specific situation, it is unnecessary). Next, to solve the equation for $x$ as a function of $t$, we invoke the MAXIMA command $\text{ode2}$ by executing the statement

Then, the solution for $x$ is

$x = \%e^{-\frac{bt}{2m}} \left( \%k1 \sin \left( \frac{4k}{m - b^2} t \right) + \%k2 \cos \left( \frac{4k}{m - b^2} t \right) \right)$
Along the way, MAXIMA has requested further information about the value of a particular quantity. This request reflects MAXIMA’s recognition that the character of the solution depends on whether the motion is underdamped, overdamped, or critically damped; by typing \texttt{p}, we selected the underdamped case. Note, incidentally, that \texttt{ode2} uses the symbols \%k1 and \%k2 for the initially undetermined integration constants in the solution to a second-order differential equation.\footnote{Note that, despite its name, \texttt{ode2} can solve both first- and second-order ODEs. The symbol \%c is used for the (single) integration constant in the solution to a first-order equation.}

Now, let’s clean up the expression a bit by introducing \(\omega\) for the frequency \(\frac{1}{\pi} \sqrt{\frac{4k}{m} - \frac{b^2}{m^2}}\) and \(\beta\) for the combination \(\frac{b}{2m}\) with the statements\footnote{The function \texttt{subst} will substitute its first argument for every occurrence of its second argument in the expression identified by the third argument. To work, the second argument must identify a subexpression that actually occurs as a part (in the technical sense; see Section 6.5) of the full MAXIMA expression. We could not, for example, substitute \(\omega\) for \(\sqrt{\frac{4k}{m} - \frac{b^2}{m^2}}\) because this latter expression does not occur as a recognizable entity. Since the argument of the trigonometric functions would be stored internally as \([/,[*,\sqrt{...},t],2]\), the quantity \([/,[*,\sqrt{...},2]\) does not appear and will not be found.}

\begin{verbatim}
(%i4) subst( 2*omega, sqrt( 4*k/m - b^2/m^2 ), %);
(%o4)
 x = %e^{-bt} \left(\%k1 \sin(\omega t) + \%k2 \cos(\omega t)\right)

(%i5) subst( 2*m*beta, b, %);
(%o5)
 x = %e^{-\beta t} \left(\%k1 \sin(\omega t) + \%k2 \cos(\omega t)\right)
\end{verbatim}

Then we impose more explicit initial conditions with the statement

\begin{verbatim}
(%i6) ic2( %, t=0, x=x[0], 'diff(x,t)=0 );
Set initial conditions \((x(0) = x_0, v(0)=0)\).
(%o6)
 x = \%e^{-\beta t} \left(\frac{x_0 \beta \sin(\omega t)}{\omega} + x_0 \cos(\omega t)\right)
\end{verbatim}

We leave it to the reader to use MAXIMA to verify that this solution in fact satisfies the original equation and the imposed initial conditions.

Like the function \texttt{solve}, \texttt{ode2} actually invokes a succession of methods in turn until it finds one that works. The system variable \texttt{method}, which can be displayed with the statement

\begin{verbatim}
(%i7) method;
Request method used.
(%o7)
 constcoeff
\end{verbatim}

is set by \texttt{ode2} to the method actually used.

The difference between the function \texttt{diff} and the function \texttt{'diff} deserves more explicit explanation. As we said above, the first of these, when invoked, will result in immediate evaluation of any derivatives that \textit{can} be evaluated while the second suppresses evaluation of derivatives until a later time when that evaluation is explicitly requested. To illustrate the difference, suppose we define the function

\begin{verbatim}
(%i8) f(y) := a * sin(omega * y);
(%o8) f(y) := a \sin(\omega y)
\end{verbatim}

Then, the two statements

\begin{verbatim}
(%i9) diff( f(y), y );
(%o9) a \omega \cos(\omega y)
\end{verbatim}

\footnote{We would prefer to substitute \(\gamma\) for \(b/2m\). Unfortunately (see footnote 8), \texttt{gamma} in MAXIMA generates \(\Gamma\), not \(\gamma\), primarily because \(\gamma\) in Maxima is reserved for the Euler-Mascheroni constant 0.577215\ldots. Evidently, that assignment cannot be turned off, even temporarily.}\footnote{We would prefer to substitute \(\gamma\) for \(b/2m\). Unfortunately (see footnote 8), \texttt{gamma} in MAXIMA generates \(\Gamma\), not \(\gamma\), primarily because \(\gamma\) in Maxima is reserved for the Euler-Mascheroni constant 0.577215\ldots. Evidently, that assignment cannot be turned off, even temporarily.
reveal the difference. In the first, the derivative is evaluated; in the second, it is left unevaluated. If, at some later time, we wish to evaluate the derivative in \( %o10 \), we could again use the function \texttt{ev}, supplying an optional argument to convey the function in \( %o10 \) that we wish \texttt{ev} to evaluate. The statement

\[
(\texttt{%i11}) \ \texttt{ev( \%o10, diff );}
\]

accomplishes this objective.

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, right after line \( %o2 \) we might have submitted the statements

\[
(\texttt{%i12}) \ \texttt{subst( omega\[0\]^2*m, k, %o2 );}
\]

\[
(\texttt{%o12}) \ \frac{d^2 x}{dt^2} + \frac{b}{\omega_0} \frac{dx}{dt} + \omega_0^2 m x = 0
\]

\[
(\texttt{%i13}) \ \texttt{subst(tau/omega\[0\], t, %);}
\]

\[
(\texttt{%o13}) \ \frac{d^2}{d(\tau^2/\omega_0^2)} x + \frac{d}{d(\tau/\omega_0)} x + \omega_0^2 m x = 0
\]

\[
(\texttt{%i14}) \ \texttt{subst(omega\[0\]*'diff(x,tau), 'diff(x,tau/omega\[0\]), %);} \quad \text{Remembering that } \omega_0^2 = k/m, \text{ where } \omega_0 \text{ is}
\]

\[
(\texttt{%o14}) \ m \left( \frac{d^2}{d(\tau^2/\omega_0^2)} x \right) + \omega_0 b \left( \frac{d}{d\tau} x \right) + \omega_0^2 m x = 0
\]

\[
(\texttt{%i15}) \ \texttt{subst(omega\[0\]*'diff(x,tau,2), 'diff(x,tau/omega\[0\],2),%);} \quad \text{Change the independent variable to } \tau = \omega_0 t.
\]

\[
(\texttt{%o15}) \ \omega_0^2 m \frac{d^2}{d\tau^2} x + \omega_0 b \frac{d}{d\tau} x + \omega_0^2 m x = 0
\]

\[
(\texttt{%i16}) \ \texttt{expand( %/(m*omega\[0\]^2) );}
\]

\[
(\texttt{%o16}) \ \frac{d^2 x}{d\tau^2} + b \frac{dx}{d\tau} + x = 0
\]

\[
(\texttt{%i17}) \ \texttt{subst(beta*m*omega\[0\], b, %);} \quad \text{Divide entire equation by } m\omega_0^2 \text{ and cancel common factors.}
\]

\[
(\texttt{%o17}) \ \frac{d^2 x}{d\tau^2} + \beta \frac{dx}{d\tau} + x = 0
\]

This process is admittedly tedious.\footnote{The parent program MACSYMA had a function \texttt{odeindeptran}, which facilitated the transformation of the independent variable in ordinary differential equations. MAXIMA has apparently not implemented that feature.} Still, at this point, we have swallowed all of the dimensional constants either into a rescaling of the time \((\tau = \omega_0 t)\) or into a single dimensionless parameter \((\beta = b/m\omega_0)\). We could, of course, rescale \(x\) to some chosen dimensionless quantity, but the equation is linear in \(x\) and we can simply interpret the equation as it stands in those terms. We would then solve the equation as before with the statement\footnote{Here, the question we answered with ‘\(n;\)’ is equivalent to the previous question whether the motion we seek is underdamped, critically damped, or overdamped. The answer \(\beta < 2\) is equivalent to the answer \(b/m\omega_0 = b/(m\sqrt{k/m}) < 2\) or \(b^2/\pi k < 4\) or \(b^2 < 4mk\). Thus, the \textit{dimensionless} inequality \(\beta < 2\) here is exactly equivalent to the \textit{dimensional} inequality \(b^2 < 4km\) that appeared earlier in this section.}
6.7. USING MAXIMA ...

(%i18) ode2( %, x, tau );
Is \((\beta - 2)(\beta + 2)\) positive, negative, or zero? n;
(%o18)  
\[
x = \%e^{-\beta \tau/2}\left(\%k1 \sin\frac{\sqrt{4 - \beta^2 \tau}}{2} + \%k2 \cos\frac{\sqrt{4 - \beta^2 \tau}}{2}\right)
\]

Finally, we impose the initial conditions with the statement

(%i19) ic2( %, tau = 0, x = x[0], 'diff(x,tau) = 0 );
(%o19)  
\[
x = \%e^{-\beta \tau/2}\left(x_0 \cos\frac{\sqrt{4 - \beta^2 \tau}}{2} - \frac{x_0 \beta \sqrt{4 - \beta^2 \sin(\sqrt{4 - \beta^2 \tau}/2)}}{\beta^2 - 4}\right)
\]

which—as we leave to the reader to show—is a dimensionless form of the dimensional result obtained at line (%o6) above.

(%i20) reset()$ kill(all)$

See item 6 in Section 6.1.

We have, of course, only mentioned a few (depends, ode2, ic2, and method) of the functions and system variables MAXIMA makes available for addressing first and second order ordinary differential equations. Important additional functions having a role in this context include atvalue, odeinlin, ic1, bc2, and ode_check. Beyond method, MAXIMA provides some information about what it has done in the variables intfactor, odeindex, and yp. Further, the package contrib_ode can be loaded to provide additional resources for solving ODEs. You should make it a point to read about these functions, option variables, system variables, and keywords in the MAXIMA manuals.

6.7.9 ... for Lists, Matrices, and Arrays

In MAXIMA, a list is "an ordered set of elements, separated by commas and enclosed in square brackets"; a matrix is "a two-dimensional, ordered set of elements"; and an array—the exposition of whose creation and manipulation we will leave to the MAXIMA manuals—is "an \(n\)-dimensional data structure". A list can be created by statements like

(%i11) lst : [ sin(x), cos(x), tan(x) ];
(%o11)  
[\sin(x), \cos(x), \tan(x)]

and its individual elements can be accessed with the notation \(\text{lst}[n]\), where \(n\)—the index of the element—is enclosed in square brackets and counts upwards from 1. The elements of a list can themselves be lists, as for example in

(%i12) lstlst : [ [1,2], [3,4] ];
(%o12)  
[[1,2], [3,4]]

and, as we have seen already, many of MAXIMA’s functions return their responses as lists or lists of lists.

A matrix, on the other hand, is not a list or a list of lists. Matrices are created, examined, and edited with statements like

(%i13) mat1 : matrix( [1,2], [3,4] );
(%o13)  
\[
\begin{bmatrix}
1 & 2 \\
3 & 4
\end{bmatrix}
\]
Examine (1,1) and (2,1) elements. Row and column indices start at 1 (not 0) and are enclosed in square brackets. Further, the (1,1) and (1,2) elements can alternatively be referred to as mat1[1][1] and mat1[1][2]. The first (second) index identifies the row (column).

Change (1,2) element. Verify change. Change it back.

Assorted quantities related to a particular matrix can also be created by invoking other MAXIMA functions as in the following examples:

Evaluate transpose of mat1. Evaluate inverse of mat1.

Load package to add function tracematrix. Evaluate trace (sum of diagonal elements) of mat1.

Evaluate determinant of mat1.

Other common manipulations include evaluating the standard matrix product and an element-by-element product, achieved with statements like

Evaluate the result of ordinary matrix multiplication.

Evaluate the result of element-by-element multiplication.

Finally (for this section anyway), MAXIMA includes functions for finding eigenvalues and eigenvectors, which we illustrate with the statements
6.7. USING MAXIMA

(%i16) spinx : matrix( [0, 1/sqrt(2), 0], 
[1/sqrt(2), 0, 1/sqrt(2)], 
[0, 1/sqrt(2), 0] );

(%o16)
\[
\begin{bmatrix}
0 & 1 & 0 \\
1 & 0 & 1 \\
0 & 1 & 0 \\
\end{bmatrix}
\]

Enter a matrix bound to the variable spinx. Multiline statements are achieved by typing the carriage return without a ‘;’ or by typing ⟨SHIFT/ENTER⟩ at the end of all lines except the last. (See item 2 in Section 6.1.)

(%i17) eigenvectors( spinx );

Find eigenvectors and eigenvalues. Note that the function eivects is a synonym for eigenvectors.

(%o17)
\[
\begin{bmatrix}
[-1, 1, 0], [1, 1, 1] \\
[1, -\sqrt{2}, 1] \\
[1, \sqrt{2}, 1] \\
[1, 0, -1] \\
\end{bmatrix}
\]

Show eigenvalues (\([-1, 1, 0]\)) and multiplicities (\([1, 1, 1]\))

Show first eigenvector

Show second eigenvector

Show third eigenvector

At the top level, the output of eigenvectors (eivects) is a list with two (sub)lists. The first (sub)list is itself a list of two (subsub)lists, the first of which lists the distinct eigenvalues—here \([-1,1,0]\)—and the second of which lists the multiplicity (degeneracy) of each eigenvalue—here \([1,1,1]\). The second (sub)list contains a (subsub)list for each distinct eigenvalue. For an eigenvalue with multiplicity (degeneracy) \(n\), the corresponding (subsub)list contains \(n\) (subsubsub)lists, whose elements are the components of the \(n\) eigenvectors belonging to the corresponding eigenvalue. Here, since each eigenvalue is non-degenerate, each (subsubsub)list contains only one (three-component) list, which explains the apparently redundant double brackets around that one eigenvector. Note that the eigenvectors as returned by eigenvectors are not (necessarily) normalized; the alternative function uniteigenvectors (or, more simply, ueivects) is basically identical to eigenvectors but it returns normalized eigenvectors.\(^{22}\)

(%i18) eigenvalues( spinx );

Find eigenvalues and multiplicities alone. Note that the function eivals is a synonym for eigenvalues.

(%o18) \([-1, 1, 0], [1, 1, 1]\)

(%i19) reset()$ kill(all)$

See item 6 in Section 6.1.

We have, of course, only mentioned a few (matrix, transpose, invert, tracematrix (in the package functs), determinant, eigenvectors, eivects, uniteigenvectors, eigenvalues, eivals) of the functions MAXIMA makes available for manipulating matrices. Important additional functions include addcol, addrow, adjoint, charpoly, coefmatrix, copypmatrix, diagmatrix, echelon, genmatrix, ident, minor, rank, row, submatrix, and zeromatrix. You should make it a point to read about these functions and any associated option variables in the MAXIMA manuals.

6.7.10 . . . for Vector Calculus

MAXIMA does not appear to have built in features or a suitable package for evaluating the standard differential operators usually symbolized by \(\nabla\) (gradient) and \(\nabla^2\) (Laplacian) applied to scalar

\(^{22}\)You may have to study the output on line (%o16) and the explanatory paragraph for awhile before the structure of this output becomes clear.
functions and $\nabla \cdot$ (divergence) and $\nabla \times$ (curl) applied to vector functions. We can, however, quickly define functions that will evaluate these quantities in, for example, Cartesian coordinates. More specifically, with $f$ and $V$ representing scalar and vector functions of three variables, the statements

\[
\text{gradcart}(f, x,y,z) := [\text{diff}(f,x), \text{diff}(f,y), \text{diff}(f,z)]$
\[
lapcart(f, x,y,z) := \text{diff}(f,x,2) + \text{diff}(f,y,2) + \text{diff}(f,z,2)$
\[
\text{divcart}(V, x,y,z) := \text{diff}(V[1],x) + \text{diff}(V[2],y) + \text{diff}(V[3],z)$
\[
\text{curlcart}(V, x,y,z) := [\text{diff}(V[3], y) - \text{diff}(V[2],z), \text{diff}(V[1], z) - \text{diff}(V[3],x), \text{diff}(V[1], y) - \text{diff}(V[2],x)]$
\]

define these four operators in Cartesian coordinates and reflect the definitions

\[
\nabla f(x,y,z) = \frac{\partial f}{\partial x} \hat{i} + \frac{\partial f}{\partial y} \hat{j} + \frac{\partial f}{\partial z} \hat{k} \tag{6.3}
\]
\[
\nabla^2 f(x,y,z) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} \tag{6.4}
\]
\[
\nabla \cdot V(x,y,z) = \frac{\partial V_1}{\partial x} + \frac{\partial V_2}{\partial y} + \frac{\partial V_3}{\partial z} \tag{6.5}
\]
\[
\nabla \times V(x,y,z) = \left( \frac{\partial V_3}{\partial y} - \frac{\partial V_2}{\partial z} \right) \hat{i} + \left( \frac{\partial V_1}{\partial z} - \frac{\partial V_3}{\partial x} \right) \hat{j} + \left( \frac{\partial V_1}{\partial y} - \frac{\partial V_2}{\partial x} \right) \hat{k} \tag{6.6}
\]

Because $f$, $V$, $x$, $y$, and $z$ in these definitions are variables, these definitions do not mandate the use of particular names for the variables involved.

For convenience, these definitions are combined in the command file—see Section 6.9—MAXvectcart.mac—see Table 6.2—available in the directory `$HEAD/maxima`.\(^{23}\) If this file is copied to your current default search directory,\(^{24}\) the statement

\(^{23}\)See the Local Guide for the value of `$HEAD` at your site.

\(^{24}\)See Section 6.16.3 for how to place this directory at the head of the reserved variable `file_search_maxima` so that the `load` command will, in fact, find the file.
6.8 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 MAXIMA. The simplest is the for/thru/step/do loop, which we have already met in Sections 6.7.2, 6.7.4, and 6.7.5. Briefly, the statements

\[
\text{(i1)} \quad \text{for } i \text{ from 1 thru 10 step 2 do } x[i] : i^2; \\
\text{end}
\]

create a 10-element list of zeros and then assign the squares of the odd integers to the odd elements of that list. In this structure, the component step can be omitted, in which case the index—here \(i\)—will be incremented by 1 at each pass through the loop.

25In more recent versions of MAXIMA, not only can a colon : be substituted for from in this statement but the step can alternatively appear before the thru, i.e., for \(i\) from 1 step 2 thru 10 ... We will adhere to the older syntax, which still works in these more recent versions.
Other available loop structures include the while loop and the unless loop, each of which requires a condition whose constant updating with each pass through the loop will ultimately change the condition to a value that will stop the loop. The general syntax of these loops is

\[
\text{for } \langle \text{variable} \rangle : \langle \text{start} \rangle \text{ step } \langle \text{increment} \rangle \text{ while } \langle \text{condition} \rangle \text{ do } \langle \text{block of statements} \rangle
\]

and

\[
\text{for } \langle \text{variable} \rangle : \langle \text{start} \rangle \text{ step } \langle \text{increment} \rangle \text{ unless } \langle \text{condition} \rangle \text{ do } \langle \text{block of statements} \rangle
\]

More specifically, these loops are illustrated in the statements

\[
\begin{align*}
\text{(i4)} & \quad x : \begin{bmatrix} 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 \end{bmatrix} \\
\text{(i5)} & \quad \text{for } i : 1 \text{ step } 2 \text{ while } i < 11 \text{ do } x[i] : i^2 \\
\text{(i6)} & \quad x; \\
\text{(i7)} & \quad \text{for } i : 1 \text{ step } 2 \text{ unless } i > 10 \text{ do } x[i] : i^3 \\
\text{(i8)} & \quad \text{block}(x[i] : i^3, i : i+2) \\
\end{align*}
\]

The specification of step is optional. If step is omitted, then the index—here \(i\)—will be incremented by 1. Further, if the block of statements includes a statement that adjusts \(i\), that adjustment will take place on top of the automatic incrementing of \(i\), e.g.,

\[
\begin{align*}
\text{(i8)} & \quad x : \begin{bmatrix} 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 \end{bmatrix} \\
\text{(i9)} & \quad \text{for } i : 1 \text{ unless } i > 9 \text{ do block}(x[i] : i^3, i : i+2) \\
\text{(i10)} & \quad x; \\
\text{(i11)} & \quad \text{block}(x[i] : i^3, i : i+2) \\
\end{align*}
\]

Note that (1) the index has been incremented by 3—not 2—so we adjust only \(x[1], x[4], \text{ and } x[7]\) and (2) groups of statements to be seen as a single composite statement are provided as arguments to the function block. The statements in a block are separated with commas.

Logical conditions appear not only to control loops but also to structure branches in a sequence of statements. As with most programming languages, MAXIMA also possesses the if/then/else construct, though the else clause can be omitted if there is in that case nothing to be done. Thus, for example, the statement

\[
\text{for } i : 1 \text{ thru } 10 \text{ do if } x[i] < 0 \text{ then } x[i] : -x[i];
\]

will replace each negative element in a list \(x\) with the corresponding positive value, the statement

\[
\text{for } i : 1 \text{ thru } 10 \text{ do if } x[i] > 10 \text{ then } x[i] : 10;
\]

will replace all values greater than 10.0 with the value 10.0, and the statement

\[
\text{if } a >= 0 \text{ then } b : a \text{ else } b : -a;
\]

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 then clause or the else clause will be separated by commas and enclosed within parentheses, as in

\[
\text{if } \texttt{condition} \text{ then ( statement, statement, ... )}
\]

\[
\text{else ( statement, statement, ... );}
\]

26The keyword from can replace the colon : in these statements.
For selection among more than two options, the statement would have the form

\[
\text{if condition then block of statements} \\
\quad \text{elseif condition then block of statements} \\
\quad \text{elseif ...} \\
\quad \text{else block of statements;}
\]

All conditions may be single or composite.

Finally, a word about composite conditions: The illustrations above made use of (single) conditions like \( i > 9 \), which will evaluate to \text{true} or \text{false} depending on the value of \( i \). Indeed, the MAXIMA function \text{is}, for example,

\[
(\%i10) \ i \ : \ 5$
(\%i11) \ [ \text{is}( \ i > 10 ), \text{is}( \ i < 10 ) ];
(\%o11) \ \text{[false, true]}
\]

allows the testing of a logical condition. The six logical operators \( =, <, >, \leq, \geq \), and \( <> \) can be used to test for equal, less than, greater than, less than or equal, greater than, and unequal.

In some situations, one may need to combine two or more conditions into a composite condition or negate a condition. MAXIMA makes available the logical operators \text{and}, \text{or}, and \text{not} to facilitate combining conditions and negating conditions. More specifically, with \( \langle C1 \rangle \) and \( \langle C2 \rangle \) individual conditions

- \( \langle \langle C1 \rangle \text{and} \langle C2 \rangle \rangle \) will be \text{true} if \( \langle C1 \rangle \) and \( \langle C2 \rangle \) are both \text{true} and \text{false} otherwise,
- \( \langle \langle C1 \rangle \text{or} \langle C2 \rangle \rangle \) will be \text{true} if either \( \langle C1 \rangle \) or \( \langle C2 \rangle \) is true or \text{both} \( \langle C1 \rangle \text{and} \langle C2 \rangle \) are true and \text{false} otherwise, and
- \( \text{not} \ (\langle C1 \rangle) \) will be true if \( \langle C1 \rangle \) is false and false if \( \langle C1 \rangle \) is true.

For example, the coding

\[
(\%i12) \ x \ : \ [ \ 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 ]$
(\%i13) \ i \ : \ 1$
(\%i14) \ \text{while} \ (i > 0 \text{and} \ i < 11) \ \text{do} \ (x[i] : i^3, \ i : i+2)\$
(\%i15) \ x; \ \%o15) \ [1,0,27,0,125,0,343,0,729,0]
\]

not only illustrates a (somewhat silly) composite condition controlling the \text{while} statement, but also recognizes that the \text{for} component can also be omitted \text{provided} a suitable incrementing statement is included in the block of statements within the body of the loop. Pure \text{while} loops and, for that matter, pure \text{unless} loops are legitimate. 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.

As is the case in most languages, loops and conditional structures can be nested, though constructing the syntax correctly can sometimes be challenging.

### 6.9 Command Files

In addition to responding to statements supplied interactively in i-lines, MAXIMA can read statements from a variety of command files. Many such files are included in the standard distribution as
examples or demonstrations and are executed by invoking the command example or the command demo with appropriate arguments.\textsuperscript{27} Other built-in command files, including packages, can be read and executed by one or another of the commands

\begin{itemize}
  \item \texttt{load( filename );} \hspace{0.5cm} This command will load files containing MAXIMA code and also files containing LISP code. Default file types for MAXIMA code (e.g., \texttt{.mac}) are listed in the option variable \texttt{file_type_maxima}; default file types for LISP code (e.g., \texttt{.lisp}) are listed in the option variable \texttt{file_type_lisp}. The search paths used are listed in \texttt{file_search_maxima} and \texttt{file_search_lisp}, respectively.\textsuperscript{28} The command \texttt{load} reads and executes the expressions in the file \texttt{silently}, i.e., does not display i- or o-lines and does not assign i and o labels along the way. Thus, a file to be loaded with \texttt{load} can contain no i- or o-labels to refer to previous lines and only those expressions that have been explicitly bound to mnemonic variables will be available after the command file has been executed.

  \item \texttt{batch( filename );} \hspace{0.5cm} This command will load files containing MAXIMA code. Default file types (e.g., \texttt{.mac}) are listed in the option variable \texttt{file_type_maxima}. The search path used is listed in \texttt{file_search_maxima}.\textsuperscript{29} Without stopping (unless the file includes requests for user input), \texttt{batch} reads and displays each i-line, computes and displays each o-line, and returns control to the MAXIMA prompt, having assigned all the labels (i- and o-line labels as well as explicitly assigned mnemonic labels) that would have been assigned had the statements been executed interactively. The MAXIMA environment available after the command file has been executed by \texttt{batch} is identical to the environment that would have existed had the statements been entered one at a time interactively.

  \item \texttt{batchload( filename );} \hspace{0.5cm} This command will load files containing MAXIMA code. Default file types (e.g., \texttt{.mac}) are listed in the option variable \texttt{file_type_maxima}. The search path used is listed in \texttt{file_search_maxima}.\textsuperscript{30} The command \texttt{batchload} reads and executes the expressions in the file \texttt{silently}, i.e., does not display i- or o-lines and does not assign i and o labels along the way. Thus, a file to be loaded with \texttt{batchload} can contain no \% signs to refer to previous lines and only those expressions that have been explicitly bound to mnemonic variables will be available after the command file has been executed.
\end{itemize}

For files having a default file type and stored in a directory in the applicable search path, neither the file type nor the path need be specified in the argument of the command and filenames can be supplied without quotation marks. Otherwise, the full path must be specified and the entire argument of the command must be enclosed in (double) quotation marks.

User-defined command files can be added by creating a text file containing the desired statements just as they would be typed interactively. In particular, no i-labels will appear in the file. For files having the appropriate default type and located in a directory in the search list defined by the option variable listed above, the file name \textit{without} the file type and \textit{without} quotation marks is sufficient as an argument. For other file types and/or directories, the full path and file name must be supplied and enclosed in double quotation marks.\textsuperscript{31}

Sometimes, creation of a command file facilitates repeated use of quantities not defined in standard MAXIMA. 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 MAXIMA; repeated typing of the correct parts of the sequence is not then necessary. We could, for example, create the file \texttt{testmultiply.mac} containing the statements listed in Table 6.3. Then, assuming that this file has been stored in a directory in MAXIMA’s search path, perhaps adjusted

\begin{footnotesize}\begin{itemize}
  \item See items 3 and 4 in Section 6.2.
  \item See Section 6.16.3 for a discussion of these two option variables and of ways to change them.
  \item See Section 6.16.3 for a discussion of this option variable and of ways to change it.
  \item See Section 6.16.3 for a discussion of this option variable and of ways to change it.
  \item Note that MAXIMA in Windows knows how to translate ‘/’ into ‘\’ for transmission to the operating system, so single (forward) slashes can be used to separate components of a path, even in the Windows operating system.
\end{itemize}\end{footnotesize}
Table 6.3: The MAXIMA command file testmultiply.mac. The structure /*...*/ allows us to include comments in the file, though the comments will not be displayed when the file is read and executed by MAXIMA.

/* Command file testmultiply.mac */

/* Enter matrix */
mat1 : matrix( [1,2], [3,4]);
/* Evaluate transpose */
mat2 : transpose(mat1);
/* Evaluate matrix product */
mat1.mat2;
/* Multiply element by element */
mat1*mat2;

as described in the previous paragraph, we would execute the statements in this file by submitting to MAXIMA one of the statements

- `batch( testmultiply );` which executes the statements in the file non-stop, displaying both the statements and the responses along the way and assigning all labels (i, o, and user-defined).

- `batchload( testmultiply );` which executes the statements in the file non-stop without displaying i-lines or o-lines and assigning only user-defined labels.

- `demo( "PathToFile/testmultiply.mac" );` which executes the statement in the file, pausing after each statement, displaying both the statements and the responses, and assigning all labels (i, o, and user-defined).\(^\text{32}\)

Were the directory containing the file not in the appropriate search path, specification of the full path would have to precede the file name.\(^\text{33}\)

Suppose, as a second example, that we frequently had two three-component lists defining two three-component vectors and that we wanted to be able easily to evaluate their cross and dot products. We might create a command file containing the lines listed in Table 6.4. Here, beyond the comments, the first three lines—actually a single statement—define the function `cross`, 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 `dot`, which takes the same arguments but returns a scalar value equal to the dot product of those two lists treated as three-component vectors. This file is named `crossdot.mac` and can be accessed in the directory `$HEAD/maxima`. Thus, the statement

`batchload("$HEAD/maxima/crossdot.mac");`

will load this file (silently), defining the two functions `dot` and `cross` for subsequent use.

Especially when statements or sets of statements are to be executed repeatedly, it may even be valuable to `translate` from MAXIMA’s language into LISP,\(^\text{34}\) since execution will then be faster. This

\(^{32}\)The option variable `file_search_demo` provides the search path for files submitted to `demo`. If modified by statements similar to those described above for modifying `file_search_maxima`, `PathToFile` can be omitted in the argument of `demo`.

\(^{33}\)See the note at the beginning of this chapter.

\(^{34}\)Behind the scenes, MAXIMA is written in LISP, so LISP is essentially the native language for the tool.
translation can be accomplished in two ways. We might, for example, load the file `crossdot.mac` with the above statement and translate each individual function with the statements

```lisp
translate( cross );
translate( dot );
```

Subsequent invocations of these functions *in the same session* will then make use of the LISP code rather than the original MAXIMA code. Alternatively, we could create a file containing the LISP code for both functions by submitting the statement\(^{36}\)

```lisp
translate_file( "Fullpath/crossdot.mac" );
```

thereby creating the file `crossdot.LISP` in the directory specified by `Fullpath`.\(^{36}\) The translator also creates a file named `crossdot.UNLISP`, which records pertinent information that may be useful if the translation is to be debugged. Once the file `crossdot.lisp` resides in the default directory and the system variable `file_search_lisp` has been edited to include the directory containing the LISP file, the statement\(^{37}\)

```lisp
load( "crossdot.lisp" );
```

will load `crossdot.lisp`, since the loader will find it before finding `crossdot.mac`, and the translated versions of the two functions will be available for use for the remainder of the session. Because the translated versions have this time been stored in a file, they can be loaded directly in later sessions as well; the translation itself need be done only once.

Even greater run-time efficiencies can be achieved if statements or sets of statements are compiled to convert MAXIMA code into the binary code directly executed by the computer. We might, for example, load the file `crossdot.mac` as above and compile each individual function with the statements

```lisp
compile( cross );
compile( dot );
```

Subsequent invocations of these functions *in the same session* will then make use of the compiled code rather than the original MAXIMA code. Alternatively, we could create a file containing the compiled code for `crossdot.mac` by submitting the statement\(^{38}\)

```lisp
compile_file( "Fullpath/crossdot.mac" );
```

thereby creating the file `crossdot.fasl` in the directory specified by `Fullpath`. The statement

```lisp
load( "Fullpath/crossdot.fasl" );
```

will then load `crossdot.fasl`, since the loader will find it before finding `crossdot.mac`) or `crossdot.lisp`, and the compiled versions of the two functions will be available for use for the remainder of the session. As with the translated versions, the compiled versions can be loaded directly in later sessions; the compilation itself need be done only once.

---

\(^{35}\)In this context, there does not appear to be a default path or a default file type.

\(^{36}\)The `.LISP` file is an ASCII file, so we can examine the LISP code by opening the file in a text editor.

\(^{37}\)The function `load` could have been included in the list earlier in this section but was not because it is a more general function. It searches first for a binary file, then for a LISP file, and finally for a file containing MAXIMA code, loading whichever it finds first. In fact, when it finds only MAXIMA code, it reverts to `batchload`.

\(^{38}\)In this context, there does not appear to be a default path or a default file type.
We conclude this section with a pointer to a way to achieve yet further efficiencies in execution. In translating or compiling the functions above, MAXIMA has had to assume that the variables might have any possible data type or structure. If only certain limited types are possible, then the number of cases with which the translated or compiled code would have to contend could be reduced. As a last refinement, we might use MAXIMA’s function `mode_declare`, described in detail in the MAXIMA manuals, to limit the cases MAXIMA should expect.

6.10 High-Resolution Plotting

High-resolution plotting in MAXIMA is provided by an external plotting package, which must be installed separately from MAXIMA. Several such packages are available, each providing a different plot format. We limit the discussion in this book to the default plot format, specifically `gnuplot` on Windows platforms and `gnuplot_pipes` on non-Windows platforms.39 MAXIMA plotting commands calculate the points to be plotted, write those points and appropriate commands to a properly constructed plot file created in the directory specified by the variable `maxima_tempdir` and by default named `maxout.gnuplot` (Windows) or `maxout.gnuplot_pipes` (non-Windows),41 and submit that file for execution by the plotting package.

We illustrate the production of high resolution graphics by returning to the damped harmonic oscillator. First, we enter the function generated by solving the damped harmonic oscillator equation at line (%o6) in Section 6.7.8 with the statement

```maxima
(%i1) soln : x[0] * exp(-beta*t) * (beta*sin(omega*t)/omega + cos(omega*t));
```

Then, recognizing that β and ω are not independent (ω² = k/m − β²) and hence that assigning values to β and ω in effect fixes k/m, we use the statements

```maxima
(%i2) ( beta:0.1, omega:1.0, x[0]:1.0 )$
```

---

39 Other packages provide plot formats `mgnuplot` and `xmaxima`, though not all are available on all platforms. The MAXIMA command `get_plot_option( plot_format );` will return the current plot format and the command `set_plot_option( [plot_format, DesiredFormat ] );` will set the plot format to the specified desired format.

40 See Section 6.16.2 for the default identity of this directory and for information about how to change it.

41 As described in Section 6.11.2, an option specified in the plotting command can direct the description of the graph to a different file in an alternative format. Produced in this way, the graph will not appear on the screen.
to assign specific values to $\beta$, $\omega$, and $x_0$. Finally, we create the graph with the statement

\[(%i3) \text{plot2d( soln, [t, 0.0, 20.0] )}\]$ 

This statement allows MAXIMA to select a range appropriate to the dependent variable for the vertical axis, plots $\text{soln}$ as a function of $t$ over the interval $0.0 \leq t \leq 20.0$,\(^{42}\) and labels the axes with default values that may or may not be what would be preferred.

By default either in the GUI or the CLI,\(^{43}\) graphs are displayed in a separate window, and display of a graph will prevent execution of further MAXIMA commands until that graph window has been closed. These graphs can be resized on the screen in the usual way by dragging the borders with the mouse.

Alternatively, but only in the GUI, graphical displays can be embedded in the wxMaxima window by prepending the characters 'wx' to the plotting function, e.g. $\text{plot2d} \rightarrow \text{wxplot2d}$, etc. In this case, once the graph has been created, additional commands can be executed without deleting the graph. Once created, an embedded graph cannot be resized, though setting the system variable $\text{wxplot\_size}$ to a two-component list, e.g.,

\[\text{wxplot\_size : [500, 250]}\]$ 

will specify in pixels the size of the region containing all subsequent plots. (The default value is $[600, 400]$.) If the size of only one plot is to be changed from the global value of $\text{wxplot\_size}$, a command of the form

\[\text{wxplot2d( ... ), wxplot\_size=[500, 250]}\]$ 

will do the job.

Assorted plot options allow more detailed customization of the resulting graph. For example, the statement

\[(%i4) \text{plot2d( soln, [t, 0.0, 20.0], [y, -2.0, 2.0] )}\]$ 

exploits the plot option $y$ to override the default range on the vertical axis,\(^{44}\) and the statement

\[(%i5) \text{plot2d( soln, [t, 0.0, 20.0], [y, -1.0, 1.0],
[ylabel, "x"], [color, black], [grid2d, true],
[label, ["beta=0.1, omega=1.0", 10.0, 0.5]] )}\]$ 

exploits the plot options $y$, title, xlabel, ylabel, color, grid2d, and label and generates a plot in which the range on both axes, the title, the labels on both axes, the color of the resulting graph, the drawing of full grid lines, and the addition of the indicated label starting horizontally at the point 10.0 and centered vertically at the point 0.5 are all specified by assigning appropriate values to several plot options.

MAXIMA also allows the presentation of more than one graph on a single set of axes. For example, we might add the velocity $v$ to the position $x$ in the graph of the previous paragraph with the statements

\(^{42}\text{The argument specifying the range of the independent variable is mandatory. In that argument, the variable name will be whatever name identifies the independent variable.}\)

\(^{43}\text{or, for that matter, in the third interface identified in the second paragraph of this chapter.}\)

\(^{44}\text{In this context, the variable $y$ is always used for that axis, regardless of the actual name of the function being plotted.}\)
Figure 6.1: Graph of position and velocity for a damped oscillator with $\beta = 0.1$, $\omega = 1.0$, and $x(0) = x_0 = 1.0$, and $v(0) = 0.0$.

which exploit the additional plot options *style* and *legend* and the (sub)plot option *lines* to arrange for different line weights for the two graphs and the inclusion of an explanatory legend. The resulting graph is shown in Fig. 6.1.

By default, the two-dimensional graphs produced above are drawn with an adaptive routine which adjusts the values at which the functions are evaluated to optimize the smoothness of the graph. The plot option *nticks*, set by including a component like `[nticks, 50]` in the *plot2d* command, sets the initial number of points used in the adaptive algorithm that decides ultimately how many points actually to use. The default value of *nticks* is 29.

We could explore other features of this motion by, for example, changing the value of $\beta$ and replotting the graph with the statements

or, returning to the original value of $\beta$, we could produce a graph of the trajectory in the phase plane ($\text{solndot}$ versus $\text{soln}$) with the statements

```plaintext
(%i18) beta : 0.2$
(%i19) plot2d( [parametric, soln, solndot, [t, 0.0, 20.0]], [x, -1.0, 1.0],
   [title, "Damped Oscillator"], [xlabel, "time"], [ylabel, "x,v"],
   [color, black], [grid2d, true], [style, [lines,4],[lines,2]],
   [legend, "x", "v"] )$
```
Figure 6.2: Graph of velocity versus position for a damped oscillator with $\beta = 0.1$, $\omega = 1.0$, and $x(0) = 1.0$, and $v(0) = 0.0$.

\begin{verbatim}
[y, -1.0, 1.0], [title, "Damped Oscillator"], [xlabel, "x"],
[ylabel, "v"], [color, black], [style, [lines,4]], [same_xy, true] )$
\end{verbatim}

In particular, this last statement sets the plot option same_xy to true to stipulate that the units on horizontal and vertical scales shall be the same for the display. The resulting graph is shown in Fig. 6.2.

Note that, in the Gnuplot window, you can generate an expanded view of a selected portion of the display by moving the cursor to one corner of the rectangle bounding the area to be expanded, clicking the right mouse button, moving the cursor to the diagonally opposite corner, and clicking the left mouse button. Clicking the left mouse button on the icon pictured as a circle with a left-pointing arrow in the window’s tool bar restores the previous display.

MAXIMA can, of course, also produce surface plots and contour plots in several different ways. The simplest and quickest surface plot is produced by statements like

\begin{verbatim}
(%i12) fct : sin( 2.0*%pi*x ) * sin( 3.0*%pi*y )$
(%i13) plot3d( fct, [x, 0.0, 1.0], [y, 0.0, 1.0], [title, "Mode 2,3"],
        [grid, 20, 20], [legend, ""] )$
\end{verbatim}

Several plot options for plot3d (and wxplot3d in the GUI), some not here illustrated, include

- grid, which specifies the number of grid points to be used along the $x$ and $y$ axis, respectively. The default value is 30, 30.
- legend, which specifies a label added to the entire graph. The default value produces the expression conveying the plotted function fct as the label. Setting the plot option to "" (as here) or to false suppresses the legend altogether.
- palette, which specifies the spectrum of colors to be used in shading each element of the surface to convey the value of fct on that element. The statement at line %i13 exploits the
default value of this option. The default value, however, can be overridden to change the choice of colors. Alternatively, including the component \[\text{palette, false}\] in the arguments of plot3d will suppress the shading altogether, producing a wire mesh diagram. Including the component \[\text{palette, [value, 0,0,0,1]}\] will use a gray scale to convey the values of \(fct\).\(^{45}\)

- **mesh_lines_color**, which controls the color of the lines defining the mesh when a palette is being used. The default value is \text{black} but other colors can be substituted. The lines can be suppressed altogether by including the component \[\text{mesh_lines_color, false}\] among the arguments of plot3d. This option is not effective when \text{palette} is false.

- **color**, which specifies the color of the mesh lines when \[\text{palette, false}\] is operative.

- **color_bar**, with default value \text{false} for plot3d, which, when set to \text{true}, produces a bar conveying the colors associated with values of the \(z\) coordinate.

Thus, the statement\(^{46}\)

\[
\text{plot3d}( fct, [x, 0.0, 1.0], [y, 0.0, 1.0], [title, "Mode 2,3"],
[grid, 20, 20], [legend, ""], [palette, false], [color, black] )$
\]

and the statement

\[
\text{plot3d}( fct, [x, 0.0, 1.0], [y, 0.0, 1.0], [title, "Mode 2,3"],
[grid, 20, 20], [legend, ""], [palette, [value, 0, 0, 0, 1]] )$
\]

will produce the graphs in Fig. 6.3. Pressing and holding \text{ML} with the cursor somewhere on the graph produced by plot3d (but not wxplot3d) and moving the cursor will rotate the diagram to facilitate fathoming its features.

An alternative display replaces the (three-dimensional) mesh surface with contour lines in a plane and is produced for the function \(fct\) by the statement

\[
\text{contour_plot}( fct, [x, 0.0, 1.0], [y, 0.0, 1.0], [legend, false],
[title, "Mode 2, 3"], [gnuplot_preamble, "set cntrparam levels 10"] )$
\]

Here, we

- suppress display of the function \(fct\) and of a chart showing the correspondence of values with colors with \[\text{legend, false}\]. The legend but not the chart would be suppressed if we had included the component \[\text{legend, ""}\] in the argument list, but the chart lies on top of—and obscures—the plot.

- override the default number of levels (3) by including the component \[\text{gnuplot_preamble, "set cntrparam levels 10"}\]. This statement, which exploits the MAXIMA plot option \text{gnuplot_preamble} communicates the \text{gnuplot} command \text{set ...} to the external plotting package, positioning that command at the beginning of the file created by \text{contour_plot} so that \text{gnuplot} will execute that instruction before generating the plot from that file. The illustrated statement specifies that 10 contour levels be drawn; the default is 5. Alternative options include

\(^{45}\)For details on constructing alternative values for \text{palette}, see the MAXIMA manuals.

\(^{46}\)To produce the graph on the left in Fig. 6.3, this statement was executed on a Linux computer by xmaxima. Executed on a Windows computer with either xmaxima or wxmaxima and on a Linux computer with wxmaxima, the statement appears to ignore the stipulation \[\text{color, black}\]. Note that Section 12.4 of the Version 5.38.1 MAXIMA manual asserts that, with plot3d, if \text{palette} is given the value \text{false}, the surfaces will not be shaded but represented with a mesh of curves only and the colors of the lines will be determined by the option \text{color}. Note also that replacing the option \[\text{color, black}\] with \[\text{gnuplot_preamble, "set lt 1 lc rgb 'black' lw 2; set lt cycle 1"}\] achieves the desired coloring. (The option \text{gnuplot_preamble} is described in the penultimate paragraph of this section (Section 6.10).)
Figure 6.3: The upper graph shows a mesh surface plot of the function $z(x, y) = \sin(2\pi x) \sin(3\pi y)$ and the lower graph shows a shaded surface plot of the function $z(x, y) = \sin(2\pi x) \sin(3\pi y)$.

- \texttt{set cntrparam levels discrete x1, x2, x3, \ldots}, which specifies the levels be drawn at the values $x1, x2, x3, \ldots$, and
- \texttt{set cntrparam levels incremental start, incr, end}, which specifies that levels be drawn at values starting with \texttt{start}, ending with \texttt{end}, and stepped in increments \texttt{incr} between \texttt{start} and \texttt{end}.

The contour plot produced by the above statement has several deficiencies, which we correct as follows:

- The resulting plot is displayed in a rectangular area, which distorts the shape of the contour lines. A better plot would use equal increments on both axes and, because the range on the two axes is the same, yield a square display. Achieving that adjustment involves invoking another option within the \texttt{set} command to \texttt{gnuplot}, expanding the above statement to be

  \begin{verbatim}
(%i17) contour_plot( fct, [x, 0.0, 1.0], [y, 0.0, 1.0], [legend, false],
                   [title, "Mode 2, 3"], [gnuplot_preamble, "set cntrparam levels 10;
  \end{verbatim}
Note that several gnuplot commands can be incorporated within the same pair of quotation marks in the gnuplot_preamble option if the commands are separated from one another with a semicolon.

- Even with this first adjustment, however, the contour lines are still cycled through a sequence of colors and the resulting display is still not particularly suited to a black and white presentation. We can force the display to use only the first color in the set through which it by default cycles by further embellishing the statement to use the keywords linetype (which can be abbreviated lt) and cycle, specifically

\[
\text{contour_plot( fct, [x, 0.0, 1.0], [y, 0.0, 1.0], [legend, false],}
\]
\[
\text{[title, "Mode 2, 3"], [gnuplot_preamble, "set cntrparam levels 10;}
\]
\[
\text{set size ratio -1; set linetype cycle 1"] )}
\]

This modification instructs gnuplot to use only the first color as it cycles through the contour lines.

- Unfortunately, the first color happens to be red. To make the contour lines black, make them a bit heavier, and make the border the same weight as the contour lines, we could add an invocation of the keywords linecolor (which can be abbreviated lc) and linewidth (which can be abbreviated lw; default 1) in the statement

\[
\text{contour_plot( fct, [x, 0.0, 1.0], [y, 0.0, 1.0], [legend, false],}
\]
\[
\text{[title, "Mode 2, 3"], [gnuplot_preamble, "set cntrparam levels 10;}
\]
\[
\text{set size ratio -1; set lt 1 lc rgb 'black' lw 2;}
\]
\[
\text{set lt cycle 1"] )}
\]

Here, apparently, we must enclose the color black in single quotation marks to avoid confusion with the double quotation marks enclosing the value of the option gnuplot_preamble.

- Since the border of the graph is a contour line at the value 0, we can make that border the same weight as the contour lines by adding a set command with the keyword border with the statement

\[
\text{contour_plot( fct, [x, 0.0, 1.0], [y, 0.0, 1.0], [legend, false],}
\]
\[
\text{[title, "Mode 2, 3"], [gnuplot_preamble, "set cntrparam levels 10;}
\]
\[
\text{set size ratio -1; set lt 1 lc rgb 'black' lw 2;}
\]
\[
\text{set border 15 lw 2; set lt cycle 1"] )}
\]

The integer after the word border uses the coding 1, 2, 4, and 8 for the bottom, left, top, and right borders, respectively (i.e., 1, 10, 100, and 1000 in binary), specify which border is to be set to the specified linewidth. To specify all four borders, the integer is set to 1 + 2 + 4 + 8 = 15, i.e., to 1111 in a binary notation.47

- Finally, we note that some of the contour lines—those in the lower and upper left and middle right—correspond to positive values of the function and others—those in the middle left and lower and upper right—correspond to negative values. To convey these differences, we invoke one more MAXIMA plot option—label—to place plus and minus signs in the appropriate regions of the contour map. The resulting statement

\[
\text{(\%i20) contour_plot( fct, [x, 0.0, 1.0], [y, 0.0, 1.0], [legend, false],}
\]
\[
\text{[title, "Mode 2, 3"], [gnuplot_preamble, "set cntrparam levels 10;}
\]
\[
\text{set size ratio -1; set lt 1 lc rgb 'black' lw 2;}
\]
\[
\text{set border 15 lw 2; set lt cycle 1"] )}
\]

Interestingly, while the above specifications of linewidths yield the contour lines and borders of the same width on screen, the translation to a PostScript file does not preserve that identity.
CHAPTER 6. INTRODUCTION TO MAXIMA

Figure 6.4: Contour map of the function $z(x, y) = \sin(2 \pi x) \sin(3 \pi y)$.

```plaintext
contour_plot( fct, [x, 0.0, 1.0], [y, 0.0, 1.0], [legend, false],
    [title, "Mode 2, 3"], [gnuplot_preamble, "set cntrparam levels 10;
    set size ratio -1; set lt 1 lc rgb 'black' lw 2;
    set border 15 lw 2; set lt cycle 1"],
    [label, ["+",0.23,0.84], ["+",0.23,0.17], ["+",0.73,0.5],
        ["-",0.23,0.5], ["-",0.73,0.17], ["-",0.73,0.84] ]);
```

achieves this objective. Note that it took a bit of experimentation to place the plus and minus signs in the center of each region.

The final graph is shown in Fig. 6.4.

(`%i21`) reset()$ kill(all)$

See item 6 in Section 6.1.

We have, of course, only mentioned a few of the functions and plot options MAXIMA makes available for generating graphical displays. Additional plot options include axes, azimuth, box, elevation, label, logx, logy, pdf_file, same_xyz, xtics, xy_scale, ytics, and yx_ratio. You should make it a point to read about these plot options in the MAXIMA manuals.

### 6.11 Making Hard Copy

The method used to make hard copy of MAXIMA output depends both on whether that output is textual or graphical and on whether a Windows or non-Windows UNIX version is in use.

#### 6.11.1 ... of Text

Though each option may work well in only some cases, text in a MAXIMA display can be printed by several routes:
• In the GUI (WX-MAXIMA) on all platforms, selecting ‘Print’ from the File menu will direct the contents of the MAXIMA workbook (text and special symbols, including Greek letters) to the printer selected in the resulting window. Embedded graphics will also be correctly printed via this procedure.

• In the CLI (X-MAXIMA) on all platforms, highlighting the text to be printed, selecting ‘copy’ from the Edit menu, pasting the text into an available text editor, and saving the result will provide a file that can then be sent to a printer by using the features of the text editor or the operating system. Greek letters cannot be displayed in or printed from the CLI. Printing graphics from this interface requires use of the commands described in Section 6.11.2.

• In the in-shell interface on some platforms, text can be cut from the MAXIMA dialog and pasted into an available text editor to create a file that can subsequently be printed. Greek letters cannot be displayed in or printed from this interface. Printing graphics from this interface requires use of the commands described in Section 6.11.2.

• In the GUI, the CLI and the in-shell interface on all platforms, all or part of the textual dialog carried out in any conversation with MAXIMA can be captured in a text file that can subsequently be directed to a printer or incorporated in other documents, perhaps with editing, though in some cases with some text editors the resulting file may be poorly formatted. We begin by opening an output file with the statement

   \texttt{writefile( "FullFileName" ); or appendfile( "FullFileName" );}

   where \texttt{FullFileName} must contain the full path to the intended destination of the file to be written. Two system-defined option variables—\texttt{maxima_tempdir} and \texttt{maxima_userdir}—store paths to possibly useful directories and may be useful in this context.\(^{48}\) Rather than explicitly typing a full path, one can construct the path within the argument of \texttt{writefile} with statements like

   \texttt{writefile( concat( maxima_tempdir, "/FileName" ) );
   writefile( concat( maxima_userdir, "/FileName" ) );

   Here, \texttt{FileName} must include the appropriate file type.\(^{49}\) Once this statement has been executed, all subsequent dialog will appear on the screen and will also be written into the specified file, with \texttt{writefile} creating a new file that overwrites any existing file of the specified name or with \texttt{appendfile} adding the new dialog at the end of an existing file.\(^{50}\) The statement

   \texttt{playback();

   which redisplay the entire dialog or, with appropriate arguments (see the MAXIMA manuals), selected parts of the dialog, can be used to avoid reentering statements already executed if we decide after the fact that we should have captured them and their output into the file. To turn off capturing into the file, use the statement

   \texttt{closefile();

   While the resulting file can be edited, copied into other files, and/or printed, it cannot be reloaded into MAXIMA.\(^{51}\)

---

\(^{48}\) See Section 6.16.2 for information about the default values of these two variable and how to change those values.

\(^{49}\) Of course, the user must have write access to the identified directory. Also look again at the syntactical issue discussed in footnote 31.

\(^{50}\) Actually, whether \texttt{writefile} overwrites or appends to an existing file may depend on the particular implementation of the user’s operating system.

\(^{51}\) To save MAXIMA dialog in a way that can be reloaded, see the \texttt{save} and \texttt{stringout} commands in the MAXIMA manuals.
• MAXIMA’s capability to write \LaTeX\ descriptions of equations on the screen and into an output file is described in Section 6.12.

6.11.2 . . . of Graphs

While embedded graphs in the GUI can be sent to the printer as described in the first bullet in Section 6.11.1, obtaining hard copy of a graph produced in a separate pop-up window on any platforms and with plot_format gnuplot in Windows and plot_format gnuplot.pipes in non-Windows requires a different approach. In these contexts, a description of the graph can be written into a file by re-executing the plot command (plot2d, plot3d, or contour.plot), adding to the optional arguments one of the following components:

- \texttt{[pdf\_file, "FileName"]} Produces output in PDF format
- \texttt{[png\_file, "FileName"]} Produces output in PNG format
- \texttt{[ps\_file, "FileName"]} Produces output in (encapsulated) PostScript format
- \texttt{[svg\_file, "FileName"]} Produces output in SVG format

where \texttt{FileName}, which should include the file type and must be enclosed in (double) quotation marks, will be specified in one of two ways:

1. If it does not contain one or more (forward) slashes, the file will be written into the directory specified by the option variable \texttt{maxima\_tempdir}.\(^{52}\)

2. If it does contain one or more (forward) slashes, the full path, including the filename itself, must be specified, and the output file will be placed in the specified location.

In either case, the user must have write access to the specified directory.

6.12 Output in \LaTeX\ Format

A particularly useful feature of MAXIMA is its ability to write \LaTeX\ versions of expressions to the screen or into a file. Full details are laid out in the MAXIMA manuals. In brief, the statement

\texttt{tex( expr );} or \texttt{tex( label );}

converts the specified expression or the expression assigned to the specified label into \LaTeX\ format and displays the result in the \texttt{MAXIMA} window. If, however, invocation of \texttt{tex} includes a destination, e.g.,

\texttt{tex( expr, "FullPathName.tex" );} or \texttt{tex( label, "FullPathName.tex" );}

where \texttt{FullPathName} includes the full path and the file type, the output will instead be written into the specified ASCII file. If the specified file does not exist, it will be created; if the file does exist, the output from this statement will be appended to the existing contents. Once created, this file can, of course, be edited or incorporated into other \LaTeX\ documents.

Two system-defined option variables—\texttt{maxima\_tempdir} and \texttt{maxima\_userdir}—store paths to possibly useful directories and may be useful in this context.\(^{53}\) Rather than explicitly typing a full path, one can construct the path within the argument of \texttt{tex} with statements like

\(^{52}\)See Section 6.16.2 for information about \texttt{maxima\_tempdir}.

\(^{53}\)See Section 6.16.2 for information about the default values of these two variable and how to change those values.
6.13 ANIMATION

Here, FileName must include the appropriate file type.\(^{54}\)

Automatic translation of arbitrary expressions into T\(\TeX\) is a daunting task. While fairly so-
plicated, the \texttt{MAXIMA} translator is not perfect. Edits will often be necessary to “correct” the
translator’s glitches. In particular, for expressions entered explicitly and expressions identified either
by a user-defined label or an o-line label, the translator uses the T\(\TeX\) characters ‘\$\$’ to begin and
end displayed equations, and we may wish to replace those with the \texttt{\LaTeX} forms \texttt{\begin{equation}}
and \texttt{\end{equation}}. We may also wish to delete the translator’s provision of an explicit equation
number equal to the \texttt{MAXIMA} label on the line. Provided we use names like \texttt{omega} and \texttt{Omega}
for variables in \texttt{MAXIMA}, the translator will convert them into \texttt{\omega} and \texttt{\Omega}. For i-line
labels, the translator simply encloses the string of ASCII characters on the i-line in a \texttt{verbatim}
environment.

6.13 Animation

Beyond the capacity for three-dimensional diagrams produced by \texttt{MAXIMA} to be dragged to dif-
ferent orientations with the cursor (see the discussion of \texttt{plot3d} towards the end of Section 6.10), a
detailed discussion of \texttt{MAXIMA}’s capabilities for producing animated displays is beyond the scope of
this book. Those interested in animations should read about the dynamics package in the MAXIMA
manuals.

6.14 Using the Workbook

Within the GUI, the window in which \texttt{MAXIMA} commands are entered has full capabilities as a
workbook.\(^{55}\) We have so far met only \texttt{Input} cells, which contain executable \texttt{MAXIMA}
statements (and possibly the associated output). These cells are bracketed along the left edge of the window with
a bracket that has at its top a line sloping from the southwest to the northeast and, when output is
also present, a short horizontal line separating input from output. For purposes of organizing larger
documents, \texttt{Text} cells, \texttt{Title} cells, \texttt{Section} cells, \texttt{Subsection} cells, and \texttt{Subsubsection} cells can contain
text in various styles.

Suppose we already have on-screen a \texttt{MAXIMA} workbook containing a few \texttt{MAXIMA} state-
ments and their output. Suppose, further, that we want to precede each statement with an explana-
tory paragraph. We could achieve that end for one of the statements by

1. Placing the cursor—a horizontal blinking line—at the point in the document where you wish
to insert some text;
2. Selecting ‘Insert ⟨desired cell type⟩ Cell’ from the \texttt{CELL} menu;\(^{56}\) and finally
3. Placing the cursor—now a vertical blinking line—in the newly created space (if it isn’t already
there) and typing the desired text.

\(^{54}\)Of course, the user must have write access to the identified directory. Also look again at the syntactical issue
discussed in footnote 31.
\(^{55}\)The counterpart in \texttt{MAXIMA} competitors is more commonly called a notebook.
\(^{56}\)Space will be opened at the point of the cursor and marked at the left edge with a symbol to convey the type of
cell created. Most of the time, a \texttt{Text} cell will be appropriate, but different levels of heading can be inserted with the
other types of cell.
This process can, of course, be repeated to provide a textual description for each of the remaining statements. Similarly, new Input cells can be created by selecting ‘Insert Input Cell’ from the CELL menu.

The CELL menu also provides tools to hide and recreate output lines. Selecting ‘Remove All Output’ shortens the workbook by deleting all output lines. Selecting ‘Evaluate All Cells’ recreates the output. Conveniently, MAXIMA workbooks can be saved by selecting ‘Save’ or ‘Save As’ from the FILE menu and migrating in the resulting browser to select the destination of the saved file, which will by default be given the file type .wxmx. In a subsequent session, the file can be reloaded most easily by selecting ‘Open’ from the FILE menu and then browsing to find and click ML on the desired file. Generally, output lines will be deleted before saving, since that output can be easily regenerated after the workbook has been reloaded into MAXIMA.

Beyond saving the workbook as a .wxmx file, MAXIMA can also export the workbook in a variety of formats. Selecting ‘Export …’ from the FILE menu brings up a browser in which one can select .html, .mac, or .tex as the type of file to be output, browse to the desired directory, and specify a suitable file name. As with most automatic creation of these files, the result is likely to be a bit cumbersome and some editing, particularly of a .tex file, may be warranted.

This short section can, of course, not do full justice to the wide spectrum of features of MAXIMA’s workbook interface. You must work with it, explore the menus in the GUI, and read the manual.

### 6.15 Pattern Matching

MAXIMA offers pattern matching as a powerful type of search and replace, as well as a means to construct our own commands. Essentially, this capability provides a means to tell MAXIMA how to recognize certain structures regardless of the specific symbols that appear in the structures, and to take some action with the recognized patterns. Pattern matching is quite complicated, so we limit the discussion here to a few simple examples. Full detail can be found in the MAXIMA manuals.

#### 6.15.1 Search and Replace: Small Angle Approximations

A human can be told something like “replace each sine term with its argument” and know what to do. Careful instruction with pattern matching can convey this same concept to MAXIMA. In this section, we develop the pattern matching commands to perform small angle approximations for sine and cosine terms.

First, we need to define a pattern variable to match “the argument” of a term. The argument can be anything, so we need a variable which will match anything. To this end, we use the command `matchdeclare`, whose syntax is

```lisp
matchdeclare( PatternVariable, Test )
```

Once this statement has been executed, `PatternVariable` will match anything for which `Test` is true.\(^{57}\) For example, the statement

```lisp
matchdeclare( x, freeof("+", "-") )
```

declares that the variable `x` will match any expression that does not contain either a “+” sign or a “-” sign, and the statement

\(^{57}\)See the MAXIMA manuals for an enumeration and description of the available tests.
matchdeclare( arg, true )

declares that the variable arg will match anything at all. After a test is performed, if the match is successful, the pattern variable has a value equal to whatever it matched.

Second, we need to use the command defrule to define a rule for carrying out replacements. The general syntax of a statement using this command is

defrule( RuleName, PatternVariable, Replacement )

Since the value of a pattern variable is whatever it matched, a pattern variable can be used in the replacement to mean the same thing that it matched in the pattern. For example, the statement

defrule( sine, sin(arg), arg )

defines a rule named sine that, when applied to an expression, will replace every appearance of sin(arg) with arg, regardless of what the specific argument of the sine function happens to be.

Third (and finally), we need the commands apply1, apply2, and applyb1 to apply the defined rules in various ways to an expression. All of these commands use a common syntax of the form

apply*( Expression, Rule1, Rule2, ..., RuleN )

They differ in the order in which the various rules are applied. apply1 applies the first rule over and over to the expression until it effects no further changes, then applies it to every subexpression until it effects no further changes, continuing to the lowest level subexpressions. The command then starts over with the second rule and follows the same process, and so on with the rest of the rules. In contrast, apply2 applies all of the rules to the expression, then moves to the subexpressions and applies all of the rules to each of them one at a time, and so on down to the lowest level subexpressions. The applyb1 works the same as the apply1, but works from the bottom up, starting with the lowest subexpression.

As a specific example, we construct rules which will do small angle approximations. We begin by defining a suitable expression with the statement

(\%i1) terms : sin(x^2) + cos(2*x) + sin(x+3);
(\%o1)   \sin(x^2) + \sin(x + 3) + \cos(2x)

Then, we invoke the statements

(\%i2) matchdeclare(arg, true)$
(\%i3) defrule(sine, sin(arg), arg);
(\%o3)   sine : sin(arg) \rightarrow arg
(\%i4) defrule(cosine, cos(arg), 1);
(\%o4)   cosine : cos(arg) \rightarrow 1
(\%i5) apply1(terms, sine, cosine);
(\%o5)   x^2 + x + 4
(\%i6) reset()$ kill(all)$

Define pattern variable arg to match anything.

Define rule to replace sin(arg) by arg.

Define rule to replace cos(arg) by 1.

Apply rules to expression.

See item 6 in Section 6.1.

to complete the process. This result is, in fact, correct, considering what we told MAXIMA to do. Here x + 3 from the sine was combined with the 1 from the cosine to yield x + 4.

This result is not, however, correct mathematically, since x + 3 is certainly not small if x is. We need to modify our pattern variable so that it does not match expressions with addition or subtraction in them (or, equivalently, to match only terms containing no addition or subtraction). To this end, we modify the items in the previous script to be
(\%i1) terms : sin(x^2) + cos(2*x) + sin(x+3)$
(\%i2) matchdeclare(arg, freeof("+", "-"))$
(\%i3) defrule(sine, sin(arg), arg)$
(\%i4) defrule(cosine, cos(arg), 1)$
(\%i5) apply1(terms, sine, cosine);
(\%o5) sin(x + 3) + x^2 + 1

(\%i16) reset()$ kill(all)$

This time, we have defined the pattern variable arg to match only those expressions that are themselves free of plus and minus signs and the result on line (\%o5) is now mathematically correct when x is small (except that we haven’t recognized the wisdom in this case of replacing \cos 2x with \(1 - (2x)^2/2\)). (Modification of the rule cosine would address that issue.)

### 6.15.2 Simplification Rules

MAXIMA has a set of rules which it applies to every term it sees in order to simplify them. For example, \(x \cdot x\) is automatically changed to \(x^2\), \(3 + 2\) to 5, \(4/6\) to \(2/3\), etc. With pattern matching, we can define additional rules to be added to MAXIMA’s automatic simplification routine. The process is similar to the definition of a rule above, except that the substitution is done without the use of an apply command. The commands to make this automatic substitution rule are tellsimp and tellsimpafter, the former defining a new rule that will be applied before the existing ones and the latter defining a new rule that will be applied after the existing ones. The syntax for both is of the form

\[
tellsimp(\ Pattern, \ Replacement ) \quad or \quad tellsimpafter(\ Pattern, \ Replacement )
\]

Names for the simplification rules are generated automatically by the system; we will never need to use them explicitly.

Suppose, for example, that we wanted small angle approximations to be applied to every term we enter into MAXIMA. We might achieve this objective with the statements

(\%i1) matchdeclare(arg, freeof("+", "-"))$
(\%i2) tellsimpafter(sin(arg), arg);
(\%o2) [sinrule2, simp-%sin]
(\%i3) tellsimpafter(cos(arg), 1);
(\%o3) [cosrule1, simp-%cos]

Once these rules have been defined, their application is automatic. Entering the expression

(\%i4) sin(x^2) + cos(2*x) + sin(x+3);

for example, yields the immediate output

(\%o4) sin(x + 3) + x^2 + 1

(\%i5) reset()$ kill(all)$

See item 6 in Section 6.1.
6.15.3 Pattern Matching Functions

Finally, we can use pattern matching to define a function to test an expression against a given pattern. This feature uses the command `defmatch` to define the new function. The syntax is

```
defmatch( FunctionName, Pattern, Parameter1, ..., ParameterN)
```

The pattern must be an expression containing only pattern variables and parameters. When the function is used, it is called with the syntax

```
FunctionName( Expression, Parameter1, ..., ParameterN)
```

If the given expression can be made to match the pattern, the function returns the values that each pattern variable and parameter matched. If MAXIMA cannot make the expression match the pattern, the function returns `false`.

As an example, we use `matchdeclare` and `defmatch` to construct a function that tests whether an expression is quadratic in a given variable. The statements

(1) `matchdeclare([a, b, c], freeof(x));`
(2) `defmatch(quadratic, a*x^2+b*x+c, x);`

define that function, and the statements

(3) `quadratic((y+2)*(y-1), y);`
(4) `[c = -2, a = 1, b = 1, x = y]`
(5) `quadratic(z^3+2*z-3, z);`
(6) `false`
(7) `quadratic((x+y)*(x+z), x);`
(8) `[c = y*z, a = 1, b = z + y, x = x]`

test that function with several examples.

6.16 Miscellaneous Occasionally Useful Tidbits

6.16.1 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. Indeed, if you wish to use backslashes, each must be doubled because a single backslash will not be properly interpreted.

6.16.2 MAXIMA’s Default Directory

Special option variables, which are defined by default when MAXIMA is launched but which can be changed by the user, are available to facilitate specifying paths for reading and writing files. Unless local adjustments have made changes (see Local Guide),
the option variable `maxima_tempdir` will point by default to a subdirectory of the user's home directory, \$UserHome/\ldots \textsuperscript{58} In particular, MAXIMA uses this directory, called the default directory or the current directory, for the storage of temporary files when plots are created.

- the option variable `maxima_userdir` will point by default to the directory \$UserHome/maxima. The default value of this directory is placed first in MAXIMA's search path.\textsuperscript{59}

(The directory \$UserHome is defined in the \textit{Local Guide}.) The current values of these two option variables can be \textit{displayed} with the statements

```
maxima_tempdir and maxima_userdir
```

and \textit{changed} with statements like

```
maxima_tempdir : "FullPathToDirectory"
maxima_userdir : "FullPathToDirectory"
```

Note, however, that changing `maxima_userdir` during a session will not automatically replace the default value at the beginning of MAXIMA's search path; the search path must also be explicitly changed.

When the default directory is unknown or ill-defined, including a full specification of a path in a quoted file name is necessary.

### 6.16.3 MAXIMA's Search Path

Whenever an explicit path is not included in the file name of a file that MAXIMA must read with \texttt{load, demo}, and some other functions, the program searches for the specified file in the sequences of directories stored in the option variables `file_search_maxima` (for MAXIMA files) and `file_search_lisp` (for lisp files). The values of these variables are set by default when MAXIMA is launched and can be \textit{displayed} with the statements

```
file_search_maxima and file_search_lisp
```

The default value of the directory\textsuperscript{60} `maxima_userdir` appears at the beginning of each of these search paths but that component in these variables must be changed \textit{explicitly} if `maxima_userdir` is changed during the session. The values of these two variables can be \textit{changed} in several ways. The statement

```
VariableName : [ "FilePath1", "FilePath2", \ldots ]
```

will redefine the variable to include \textit{only} the specified list. Alternatively, the statement

```
VariableName : append( VariableName, [ "FilePath1", "FilePath2", \ldots ] )
```

will add the specified list to the \textit{end} of the current list, and the statement

```
VariableName : append( [ "FilePath1", "FilePath2", \ldots ], VariableName )
```

\textsuperscript{58}On the author's Windows machine, `maxima_tempdir` points to \$UserHome/AppData/Local/Temp.

\textsuperscript{59}See Section 6.16.3.

\textsuperscript{60}See Section 6.16.2.
will add the specified list at the beginning of the current list. Note that, in the syntax of all of these statements, the square brackets are necessary, even if the enclosed list contains only one element.

In specifying FilePath, the wild card ### can be used to stand for the file name that is the search target and enclosing components in {...} will allow abbreviating the path. For example, if we execute the statement

```
file_search_maxima : ["c:/users/cookd/###.{mac,mc}"];
```

and then execute the statement

```
load( testfile )
```

the files c:/users/cookd/testfile.mac and c:/users/cookd/testfile.mc, should they exist, will both satisfy the search.

### 6.16.4 Customizing MAXIMA

When MAXIMA is started up, its configuration is defined by a number of default specifications built into the program. MAXIMA also responds to a user-specified initialization file, which provides a means for further automatic customization of MAXIMA in ways different from the vendor-supplied defaults. This file is, in the first instance, written initially in MAXIMA code and is named maxima-init.mac but—see Section 6.9—can be translated or compiled to speed execution. Whether in MAXIMA code, translated, or compiled, this file resides in a subdirectory of the user’s home directory, which will be reported by MAXIMA in response to the statement

```
maxima_userdir;
```

typically $UserHome/maxima. The file maxima-init.mac may contain any MAXIMA statements and is executed automatically immediately when MAXIMA is launched.

### 6.16.5 Restoring MAXIMA’s Initial State

See item 6 in Section 6.1.

### 6.16.6 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

```
(%i48) plot2d( sin(x), [x, 0.0, 10.0], [label, ["New Graph",5.0.0.0]] );
```

This statement invokes the plot option label to specify that the text ‘New Graph’ be added to the graph with the lower left corner of the text located at the point [5.0,0.0] and the text displayed horizontally. The command plot2d does not appear to provide control over the positioning of the text relative to the specified point or the orientation of the text, though the command draw2d in the draw package does offer that additional control through the options label_alignment and label_orientation. Full detail on these commands and their extension to three-dimensional graphs can be found in the MAXIMA manuals.
CHAPTER 6. INTRODUCTION TO MAXIMA

6.16.7 Space Curves

Sometimes it is desirable to view the trajectory of a particle in three-dimensional space. MAXIMA 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{plot3d} 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

$$
x(t) = \cos t ; \quad y(t) = \sin t ; \quad z(t) = \alpha t
$$

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}
(\%i1) alpha : 0.5$
(\%i2) plot3d( [cos(t), sin(t), alpha*t], [t,0,30], [y, -1,1], [z,0,20],
[same_xy,true], [xtics,-1,0.5,1], [ytics, -1, 0.5, 1],
[grid,500,1], [legend,false] );
\end{verbatim}

will produce the graph in Fig. 6.5. Here, in particular, the plot option \texttt{grid} defines the number of points at which the trajectory is evaluated in the interval on the parameter $t$. (The role of the second integer in that option is unclear.) and the options \texttt{xtics} and \texttt{ytics} stipulate that tics are to be drawn and labeled at increments of 0.5 in the interval $-1.0 \leq x, y \leq 1.0$.

6.17 Miscellaneous Commands

Several miscellaneous commands to MAXIMA are also of general utility. We should in particular be aware of at least the following commands:

- \texttt{apply()}; which facilitates application of a function to every element in a list, e.g. \texttt{apply("*",[1,2,3])}, which returns the value 6.
• `apropos()`, which returns a list of valid arguments for `describe()`, where those arguments all contain the string in the argument of `apropos()`. The statement `apropos("sin")`, for example, returns a list of all functions, option variables, and other entities containing the string ‘sin’ for which `describe()`—see Section 6.2—will give additional information.

• `assume()`, which attaches properties to symbolic variables, e.g., `assume( b > 0 )`. Assumed properties can be removed with `forget()`.

• `declare()`, which permits attaching a variety of features to variables. The statement `declare(n, integer)`, for example, tells MAXIMA to consider only integer values for the variable n. Declared features can be removed with `remove()`. The statement `features;` will result in a display of the system variable `features`, which contains a list of all the features that can be declared.

• `dispfun( FunctionName );`, which displays the definition of a user-declared function.

• `forget();`, which instructs MAXIMA to “forget” properties assigned with `assume`.

• `functions;`, which displays the names of all user-declared functions.

• `infolists;`, which displays a list of the names of the system variables that store lists of information maintained by the system.

• `labels( symbol );`, which displays a list of the i-line, o-line, and intermediate expression labels to which values have been bound.

• `loadfile( FileName );`, which restores MAXIMA to the state it had when the “loaded” file was created with the command `save`.

• `map();`, which specifies application of a given function separately to each term of an expression.

• `product();`, which multiplies given values over a specified range.

• `remfunction( FunctionName );`, which “un”defines a previously defined function.

• `remove();`, which causes MAXIMA to “forget” a feature attached to a variable by `declare`.

• `remvalue( VariableName );`, which removes MAXIMA’s knowledge of the value previously bound to a variable.

• `save( FileName, all );`, which saves MAXIMA’s current state in a file that can subsequently be read into MAXIMA to restore that state. (See `loadfile`.)

• `stringout();`, which outputs selected quantities to a specified file.

• `sum();`, which sums given values over a specified range of indices.

• `values;`, which displays the names of variables to which values have been bound.

Additional information about these commands can be displayed by entering the command name in the search box of the index in the Maxima manual.

### 6.18 References

In this chapter, we have introduced only the most important features of MAXIMA. A full description of all commands and features is contained in the MAXIMA manuals, including

• several documents and resources as described in Section 6.2.
• a number of documents, links to which can be found by entering “Maxima help” in the search window of your browser. In particular, you should find links to
  - the official MAXIMA manual at
    http://maxima.sourceforge.net/docs/manual/maxima.html
  - and
  - many other official and unofficial documents.
• more specific documents on particular MAXIMA commands, links to which can be found by entering “CommandName in MAXIMA”, e.g., “integrate in MAXIMA” in the search window of your browser.

The user is urged to browse in these manuals, being alert not only to the available functions but also to the assorted option variables and keywords that modify their behavior.

6.19 Exercises

6.1. Read the MAXIMA manuals to find out how the commands combine, distrib, expand, multthru, and xthru 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

(a) \( \frac{x}{a+b} + \frac{y}{a+b} \)  
(b) \( (a+b)(c+d) + (a+b)c \)

(c) \( (x+y)(z+w) \)  
(d) \( a(b+c(d+e)) \)

(e) \( \frac{x}{a+b} + \frac{y}{c+d} \)  
(f) \( \frac{(a-x)^2}{(a^2-2ax+x^2)^{3/2}} \)

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.

6.2. Use MAXIMA to convert each of the expressions in the left-hand column in the table below into the expression in the associated right-hand column:

(a) \( \frac{(a-x)^2}{(a^2-2ax+x^2)^{3/2}} \)  \( \Rightarrow \) \( \frac{1}{|x-a|} \)

(b) \( \sinh( \ln(x+\sqrt{x^2+a^2}) - \ln(a) ) \)  \( \Rightarrow \) \( \frac{x}{a} \)

(c) \( \frac{1}{x+\sqrt{y}} \)  \( \Rightarrow \) \( \frac{x-\sqrt{y}}{x^2-y} \)

These samples are chosen to illustrate particularly the use of radcan and ratsimp but other commands will surely also be needed. 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. Hint: For part (c), you might find it useful to look up the option variable algebraic.

6.3. Use MAXIMA to convert each of the expressions in the left-hand column in the table below into the expression in the associated right-hand column:
6.19. EXERCISES

[(a) \( \frac{d}{dx} \left( x^2 e^{-x^2} \right) \) \[ \implies -2x(x^2 - 1)e^{-x^2} \]]

[(b) \( \sin(\sqrt{a^2(a + 3x) + x^2(3a + x) + y}) \) \[ \implies \sin((a + x)^{3/2} + y) \]]

[(c) \( cg + cf + b^2d + 2abd + a^2d + b^2c + 2abc + a^2c \) \[ \implies (a + b)^2(c + d) + c(f + g) \]]

[(d) \( ae^{(b+i\omega)t} + a e^{(b-i\omega)t} \) \[ \implies a e^{-b t} (e^{-i\omega t} + e^{i\omega t}) \]]

[(e) \( ae^{(b+i\omega)t} + a e^{(b-i\omega)t} \) \[ \implies 2 a e^{-b t} \cos(\omega t) \]]

[(f) \( x^2 + y^2 + z^2 - 2a(x + y) + 2a^2 \) \[ \implies (x - a)^2 + (y - a)^2 + z^2 \]]

In several cases, you may need to invoke part, pickapart, subst, and/or substpart 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.

6.4. In Section 6.7.8, 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

\[
\frac{d^2x}{dt^2} + 2a \frac{dx}{dt} + x = 0 \quad ; \quad x(0) = 1 \quad ; \quad \frac{dx}{dt} \bigg|_{t=0} = 0
\]

of the equation deduced towards the end of Section 6.7.8, 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 \( v(0) = 0 \).

6.5. 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

\[ g(x,t) = \frac{1}{\sqrt{1 - 2xt + t^2}} = \sum_{n=0}^{\infty} P_n(x) t^n \]

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

\[
\begin{align*}
P_0(x) &= 1 & P_3(x) &= \frac{1}{7} (5x^3 - 3x) \\
P_1(x) &= x & P_4(x) &= \frac{1}{8} (35x^4 - 30x^2 + 3) \\
P_2(x) &= \frac{1}{2} (3x^2 - 1) & P_5(x) &= \frac{1}{8} (63x^5 - 70x^3 + 15x) \\
\end{align*}
\]

a. Use the generating function and MAXIMA’s capabilities for evaluating Taylor series to find the first half-dozen Legendre polynomials, extracting each as an expression bound to a variable.

\[ \text{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 \( \text{twice a constant, so } \alpha = b/2m\omega_0 \).} \]
b. Start by binding the value 1 to $P[0]$ and the value $x$ to $P[1]$ and then, using the recursion relationship, find the next several Legendre polynomials. *Hint:* The MAXIMA statements would be

\begin{verbatim}
(%i1) P[0] : 1;
(%i2) P[1] : x;
\end{verbatim}

and then

\begin{verbatim}
(%i3) P[n] := (2*n-1)*x*P[n-1]/n - (n-1)*P[n-2]/n;
\end{verbatim}

(Verify the expression on the right by using MAXIMA 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 a function $P[n]$ involving $n$ 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

\begin{verbatim}
(%i4) P[2];
(%i5) P[3];
\end{verbatim}

(Some simplifications may be necessary to cast the results in the standard form of the above table.)

c. Find the first half-dozen Legendre polynomials by using the command *diff* 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 MAXIMA (see manuals for the structure of a *for* loop), 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.)

e. Within MAXIMA, 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. MAXIMA actually knows quite a bit about many of the important special functions of mathematical physics. In particular, the package *orthopoly*, which can be loaded with the command *load( orthopoly )* and which is described fully in the MAXIMA manuals, provides two different ways to evaluate a very large number of these special functions. Take a look at that documentation and use the function *legendre_p* (syntax *legendre_p(n, x)*) in that package to determine the first several Legendre polynomials.

6.6. Figure 6.6 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 MAXIMA, (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.)
6.7. 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

\[
\begin{align*}
\psi(0^-) &= \psi(0^+) & \frac{d\psi}{dx}\bigg|_{0^-} &= \frac{d\psi}{dx}\bigg|_{0^+} \\
\psi(w^-) &= \psi(w^+) & \frac{d\psi}{dx}\bigg|_{w^-} &= \frac{d\psi}{dx}\bigg|_{w^+}
\end{align*}
\]

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 MAXIMA’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

\[
R = \left| \frac{B}{A} \right|^2 = \frac{(k^2 + \kappa^2)^2 \sinh^2 \kappa w}{4\kappa^2k^2 + (k^2 + \kappa^2)^2 \sinh^2 \kappa w} ; \\
T = \left| \frac{C}{A} \right|^2 = \frac{4k^2\kappa^2}{4\kappa^2k^2 + (k^2 + \kappa^2)^2 \sinh^2 \kappa w}
\]

Here, the vertical bars symbolize the absolute value of the complex number enclosed by them.

(d) Verify that \( R + T = 1 \).

6.8. The distribution of wavelengths \( \lambda \) in the blackbody spectrum at (absolute) temperature \( T \) is given by

\[
u(\lambda, T) = \frac{8\pi c h}{\lambda^5} \frac{1}{e^{c h / \lambda k T} - 1}
\]

where \( c \) is the speed of light, \( h \) is Planck’s constant, and \( k \) is Boltzmann’s constant. In terms of the variable \( y = ch/\lambda k T \), 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 MAXIMA, (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.

6.9. Use MAXIMA to verify (a) that the solution presented as d6 in Section 6.7.8 satisfies the differential equation and the initial conditions presented in c2 and c6 of that section and (b) that the solutions given in d6 and d17 agree with each other.

6.10. 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
\[
\begin{align*}
L_0(x) &= 1 \\
L_1(x) &= 1 - x \\
L_2(x) &= 1 - 2x + \frac{1}{2}x^2 \\
L_3(x) &= 1 - 3x + \frac{3}{2}x^2 - \frac{1}{2}x^3 \\
L_4(x) &= 1 - 4x + 3x^2 - \frac{3}{2}x^3 + \frac{1}{4}x^4 \\
L_5(x) &= 1 - 5x + 5x^2 - \frac{5}{2}x^3 + \frac{5}{24}x^4 - \frac{1}{120}x^5
\end{align*}
\]
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 MAXIMA statements might be
\[
\begin{align*}
(\%11) \quad & L[0] : 1; \\
(\%12) \quad & L[1] : 1-x;
\end{align*}
\]
and then
\[
(\%13) \quad L[n] := (2*n-1-x)*L[n-1]/n - (n-1)*L[n-2]/n;
\]
(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 a function \( L[n] \) 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 ... with statements like
\[
\begin{align*}
(\%14) \quad & L[2]; \\
(\%15) \quad & L[3]; \\
& \ldots
\end{align*}
\]
6.19. EXERCISES

(Some simplifications may be necessary to cast the results in the standard form of the above table.)

c. Find the first half-dozen Laguerre polynomials by using the command \texttt{diff} to evaluate Rodrigues’ formula.

d. Be clever and, using either matrices or loops constructed in MAXIMA (see manuals for the structure of a \texttt{for} loop), 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.)

e. MAXIMA actually knows quite a bit about many of the important special functions of mathematical physics. In particular, the package \texttt{orthopoly}, which can be loaded with the command \texttt{load(orthopoly)} and which is described fully in the MAXIMA manuals, provides two different ways to evaluate a very large number of these special functions. Take a look at that documentation and use the function \texttt{laguerre} (syntax \texttt{laguerre(n, x)}) in that package to determine the first several Laguerre polynomials.

6.11. Find the eigenvalues and eigenvectors of the matrix

\[
\begin{pmatrix}
0 & 1 & 0 \\
1 & 0 & 1 \\
0 & 1 & 0
\end{pmatrix}
\]

6.12. (a) 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{part} to explore the structure of the entity returned by the function \texttt{eigenvectors} used in part (a).

6.13. 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 \(\langle nlm|eF|nl'm'\rangle\), 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

\[
3e a_0 F \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -9/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]
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. **Hint:** Note that the command `zeromatrix` may be useful in speeding the creation of these matrices, many of whose elements are zero and the command `radcan` may be useful in simplifying results.

6.14. Develop a procedure using pattern matching that will convert a polynomial in \( x^2 \) to a polynomial in \( y \), where \( y = x^2 \).

6.15. As shown in Fig. 6.7, 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 MAXIMA 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 `diff(L, diff(x,t))` are perfectly acceptable to MAXIMA, provided the dependence of \( x \) on \( t \) is indicated, either explicitly \( x(t) \) or through a prior execution of the statement `depends(x, t)`.

6.16. In the Cartesian coordinate system illustrated in Fig. 6.8, 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

\[
KE = \frac{1}{2} m_1 \dot{x}_1^2 + \frac{1}{2} m_1 \dot{y}_1^2 + \frac{1}{2} m_2 \dot{x}_2^2 + \frac{1}{2} m_2 \dot{y}_2^2; \quad PE = m_1 g y_1 + m_2 g y_2
\]
use MAXIMA 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.

6.17. 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 MAXIMA to evaluate these integrals to $O(k^6)$ by expanding the integrands in Taylor series before evaluating the integrals.

6.18. Find the Laplace transform of each of the functions

(a) $f(t) = t^n$  (b) $f(t) = te^{-at}$  (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 MAXIMA’s `ilt` routine by first invoking the `partfrac` routine to expand the desired function in partial fractions.

6.19. 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.

6.20. 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

$$\int_{x_1}^{x_3} y(x) \, dx = \frac{\delta x}{3} \left( y_1 + 4y_2 + y_3 \right)$$

for (approximate) numerical integration.
6.21. The midpoint rule $M$ and Simpson’s rule $S$ for evaluating integrals numerically start with the assumptions that

$$\int_a^b f(x) \, dx \approx M = (b - a) f \left( \frac{a + b}{2} \right); \quad \int_a^b f(x) \, dx \approx S = \frac{b - a}{6} \left[ f(a) + 4f \left( \frac{a + b}{2} \right) + f(b) \right]$$

respectively. To deduce the first of these expressions, we approximate $f(x)$ over the interval $a \leq x \leq b$ with a constant equal to the value of $f(x)$ at the midpoint of the interval while deducing the second entails approximating the function with a quadratic polynomial that passes through $f(x)$ at the endpoints and the midpoint of the interval. The midpoint rule will clearly be 100% accurate if $f(x)$ is a constant and Simpson’s rule will be 100% accurate if $f(x)$ is a quadratic polynomial. Use symbolic manipulation to show that, surprisingly, the midpoint rule is in fact 100% accurate for the linear function $f(x) = mx + p$ and Simpson’s rule is 100% accurate for the cubic polynomial $f(x) = cx^3 + dx^2 + ex + f!$ Optional. Try to construct geometric arguments that would provide insight into the correctness of these analytic results.

6.22. Using MAXIMA, obtain graphs of the potential

$$U(x) = -\frac{U_0a^2(a^2 + x^2)}{8a^4 + x^4}$$

and the associated force $F = -dU/dx$. 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 $\tau = x/a$, i.e., rewrite the function in the form

$$\frac{U(x)}{U_0} = \frac{1 + \tau^2}{8 + \tau^4}$$

You should be able to find a similar dimensionless version of the expression giving the force.

6.23. Using MAXIMA, find complete (symbolic) solutions to each of the following problems, and use MAXIMA also to verify that the solutions you obtain actually do satisfy the original ODE and initial conditions.

a. $\frac{d^2x}{dt^2} = a$, $x(0) = x_0$, $v(0) = v_0$, 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. $m \frac{d^2x}{dt^2} = -eE_0 \cos(\omega t + \theta)$, $x(0) = x_0$, $v(0) = v_0$, 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. $m \frac{d^2x}{dt^2} = -mg + b \left( \frac{dx}{dt} \right)^2$, $x(0) = 0$, $v(0) = 0$, 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

$$m \frac{d^2x}{dt^2} = -b \frac{dx}{dt} \quad \text{and} \quad m \frac{d^2z}{dt^2} = -mg - b \frac{dz}{dt}$$

for the motion of a projectile moving under gravity in a viscous medium when the motion starts at the origin with the initial velocity $v = v_{i0} \hat{i} + v_{i0} \hat{k}$.

6.24. Using the functions \texttt{gradcart}, \texttt{lapcart}, \texttt{divcart}, and \texttt{curlcart} defined in the file \texttt{MAXvectcart.mac} described in Section 6.7.10.

a. Evaluate the gradient of (i.e., the negative of the force field associated with) each of the functions

$$V_1(x, y, z) = \frac{1}{(x^2 + y^2 + z^2)^{1/2}} \quad \text{and} \quad V_2(x, y, z) = \frac{z}{(x^2 + y^2 + z^2)^{3/2}}$$
6.19. EXERCISES

\begin{align*}
V_4(x, y, z) &= \frac{2z^2 - x^2 - y^2}{(x^2 + y^2 + z^2)^{5/2}} & V_4(x, y, z) &= \frac{e^{-a(x^2 + y^2 + z^2)^{1/2}}}{(x^2 + y^2 + z^2)^{1/2}}
\end{align*}

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 \(\mathbf{r} = x \mathbf{i} + y \mathbf{j} + z \mathbf{k}\), 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 \(V_4(x, y, z)\), \(\nabla^2 V_4(x, y, z)\).

6.25. Paralleling your development after Section 6.7.10, develop the command file \texttt{MAVvectorcyl.mac} to provide functions \texttt{gradcyl}, \texttt{lapcyl}, \texttt{divcyl}, and \texttt{curlcyl} to return the derivatives

\begin{align*}
\nabla f(r, \phi, z) &= \frac{\partial f}{\partial r} \hat{r} + \frac{1}{r} \frac{\partial f}{\partial \phi} \hat{\phi} + \frac{\partial f}{\partial z} \hat{z} \\
\nabla^2 f(r, \phi, z) &= \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial f}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 f}{\partial \phi^2} + \frac{\partial^2 f}{\partial z^2} \\
\nabla \cdot \mathbf{V}(r, \phi, z) &= \frac{1}{r} \frac{\partial}{\partial r} (r \partial f/\partial r) + \frac{1}{r} \partial Q_r/\partial \phi + \frac{1}{r} \partial Q_r/\partial z \\
\nabla \times \mathbf{Q}(r, \phi, z) &= \left( \frac{1}{r} \frac{\partial Q_z}{\partial \phi} - \frac{\partial Q_\phi}{\partial z} \right) \hat{r} + \left( \frac{\partial Q_z}{\partial r} - \frac{\partial Q_r}{\partial \phi} \right) \hat{\phi} + \frac{1}{r} \left( \frac{\partial (r \partial Q_\phi)}{\partial r} - \frac{\partial Q_r}{\partial \phi} \right) \hat{z}
\end{align*}

for evaluating these entities in cylindrical coordinates. Here, \(r, \phi, \) and \(z\) are the (cylindrical) radial, azimuthal, and axial coordinates. Then,

a. Evaluate the gradient of (i.e., the negative of the force field associated with) each of the functions

\begin{align*}
V_1(r, \phi, z) &= \frac{1}{(r^2 + z^2)^{1/2}} & 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)^{5/2}} & V_4(r, \phi, z) &= \frac{e^{-a(r^2 + z^2)^{1/2}}}{(r^2 + z^2)^{1/2}}
\end{align*}

of the cylindrical variables.

b. Verify that each field obtained in part (a) is conservative by showing that the curl of each is zero.

c. With \(\mathbf{r} = r \hat{r} + z \hat{z}\), 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 \(V_4(r, \phi, z)\), \(\nabla^2 V_4(r, \phi, z)\).

6.26. Paralleling your development after Section 6.7.10, develop the command file \texttt{MAVvecsph.mac} to provide functions \texttt{gradsph}, \texttt{lappsh}, \texttt{divsph}, and \texttt{curlsph} to return the derivatives

\begin{align*}
\nabla f(r, \theta, \phi) &= \frac{\partial f}{\partial r} \hat{r} + \frac{1}{r} \frac{\partial f}{\partial \theta} \hat{\theta} + \frac{1}{r \sin \theta} \frac{\partial f}{\partial \phi} \hat{\phi} \\
\nabla^2 f(r, \theta, \phi) &= \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial f}{\partial r} \right) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial f}{\partial \theta} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 f}{\partial \phi^2} \\
\nabla \cdot \mathbf{V}(r, \theta, \phi) &= \frac{1}{r} \frac{\partial}{\partial r} (r^2 \partial f/\partial r) + \frac{1}{r} \partial Q_r/\partial \theta + \frac{1}{r} \partial Q_r/\partial \phi \\
\nabla \times \mathbf{Q}(r, \theta, \phi) &= \frac{1}{r \sin \theta} \left( \frac{\partial (\sin \theta Q_\phi)}{\partial \theta} - \frac{\partial Q_\theta}{\partial \phi} \right) \hat{r} + \left( \frac{1}{r \sin \theta} \frac{\partial Q_\theta}{\partial \phi} - \frac{1}{r} \frac{\partial Q_r}{\partial r} \right) \hat{\theta}
\end{align*}
\[ + \frac{1}{r} \left( \frac{\partial (rQ_\theta)}{\partial r} - \frac{\partial Q_r}{\partial \theta} \right) \hat{\phi} \]

for evaluating these entities in spherical coordinates. Here, \( r, \theta, \) and \( \phi \) are the (spherical) radial, polar, and azimuthal coordinates. Then,

a. Evaluate the gradient of (i.e., the negative of the force field associated with) each of the functions

\[ V_1(r, \theta, \phi) = \frac{1}{r} \quad V_2(r, \theta, \phi) = \frac{z}{r^3} \quad V_3(r, \theta, \phi) = \frac{r^2(3 \cos^2 \theta - 1)}{r^3} \quad V_4(r, \theta, \phi) = \frac{e^{-ar^{1/2}}}{r} \]

of the spherical variables.

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 \( V_4(r, \theta, \phi) \), \( \nabla^2 V_4(r, \theta, \phi) \).

6.27. Create four three-component vectors \( \mathbf{A}, \mathbf{B}, \) and \( \mathbf{C} \) with statements like

\[
\mathbf{A} : \begin{bmatrix} A[1], & A[2], & A[3] \end{bmatrix} \quad \mathbf{B} : \begin{bmatrix} B[1], & B[2], & B[3] \end{bmatrix} \\
\mathbf{C} : \begin{bmatrix} C[1], & C[2], & C[3] \end{bmatrix} \quad \mathbf{D} : \begin{bmatrix} D[1], & D[2], & D[3] \end{bmatrix}
\]

Then, using the file `crossdot.mac` as described in Section 6.9, 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. \( (\mathbf{A} \times \mathbf{B}) \cdot (\mathbf{C} \times \mathbf{D}) = (\mathbf{A} \cdot \mathbf{C})(\mathbf{B} \cdot \mathbf{D}) - (\mathbf{A} \cdot \mathbf{D})(\mathbf{B} \cdot \mathbf{C}) \)
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 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.

2Some 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 $). Thus, in the pseudocode we are defining, the names x, position, and field identify memory cells for floating point values, the names I%, count%, and lower_limit% are valid names for cells storing integers, the names word$ and first_name$ are valid names for cells storing strings, and item@ identifies a quantity of any type.

The variables we have discussed so far are scalar or 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 array is among the 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., data, values%, or names$, depending on the data type of the elements of the array. Individual elements are identified by attaching an integer index or indices to the array name, e.g., data(I%,J%), values%(4%), or names$(12%). All by itself, the name of a scalar is sufficient to tell the compiler how much memory to reserve to store the scalar value. For 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

\[
\text{DIMENSION data(4%,251%), values%(8%), names$(25%)}
\]

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 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 \text{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 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:

  \[
  \langle \text{variable} \rangle \leftarrow \langle \text{expression determining value to be assigned} \rangle
  \]

  e.g.,\(^4\)

  \[
  \text{val} \leftarrow \sin(\pi \times x)
  \]

  \[
  \text{count}\% \leftarrow \text{count}\% + 1\%
  \]

  \[
  \text{first\_name}$ \leftarrow "\text{David}"
  \]

  etc. Any valid arithmetic expression utilizing the symbols + for addition (or string concatenation), - for subtraction, * for multiplication, / for division, and ^ 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 (\(\sin\), \(\cos\), \(\tan\), \(\sqrt{\cdot}\), \(\exp\), ...) 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:

  \[
  \text{WRITE "The value of the integral is }, \text{value}, "."
  \]

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, ... 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.

\(^3\)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.

\(^4\)We declare that integer \textit{constants} shall be explicitly identified with a trailing percent sign and that string \textit{constants} shall be enclosed in double quotation marks. Some languages use single quotation marks, while others will accept either (provided they are paired).
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

\[\text{EQUAL? } \leftarrow N@ = M@\]

in response to which the computer will assess whether \(N@\) is equal to \(M@\) 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.\footnote{We have used $\leftarrow$ and $. In later sections, we will identify the symbols used in other languages.}

\section*{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

\begin{verbatim}
val ← sin(pi*x); count% ← count% + 1%; first_name$ ← "David"
\end{verbatim}

- to spread a single statement over more than one physical line. In our generic code, we shall use the symbol $\rightarrow$ at the end of a line to indicate that the statement continues on the next line. Thus, for example, the two-line statement

\begin{verbatim}
U(I%,J%) ← 0.25 * ( U(I%+1%, J%) + U(I%-1%, J%) + U(I%, J%+1%) + U(I%, J%-1%) )
\end{verbatim}

is to be seen as a single statement.\footnote{Remember that spaces not in quoted strings are ignored by the compiler.}

- 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{verbatim}
BEGIN_BLOCK
  val ← sin(pi*x)
  count% ← count% + 1%
  first_name$ ← "David"
END_BLOCK
\end{verbatim}

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.

\section*{9.1.5 Essential Control Structures}

Four main control structures\footnote{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.} 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.

$\text{...}$

$\text{...}$
9.1. COMPONENTS OF A PROGRAMMING LANGUAGE

- Two-way selection, expressed generically with a statement like

```plaintext
IF ⟨condition⟩
  THEN ⟨block1 of statements⟩ ! the THEN clause
  ELSE ⟨block2 of statements⟩ ! the ELSE clause
END_IF
```

Here ⟨block1 of statements⟩ is executed when ⟨condition⟩ is true and ⟨block2 of statements⟩ is executed when ⟨condition⟩ is false. The ELSE clause may be omitted altogether if ⟨block2 of statements⟩ is null. The flow diagram in Fig. 9.1(a) is often used to convey this structure.

- Multi-way selection

```plaintext
CASE
  OF ⟨condition1⟩ DO ⟨block1 of statements⟩
  OF ⟨condition2⟩ DO ⟨block2 of statements⟩
  OF ⟨condition3⟩ DO ⟨block3 of statements⟩
  ...
  OF OTHERS DO ⟨blockO of statements⟩
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. The OTHERS clause may be omitted altogether if ⟨blockO of statements⟩ is null. The flow diagram in Fig. 9.1(b) is often used to convey this structure.

- Repetition

```plaintext
LOOP
  ⟨block1 of statements⟩
  EXIT_LOOP WHEN ⟨condition⟩
  ⟨block2 of statements⟩
END_LOOP
```

Here, the statements in the body of the loop are executed repeatedly until ⟨condition⟩, which is tested at the indicated point in the loop, becomes true. Looping will continue forever unless repeated execution of the statements eventually causes ⟨condition⟩ 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

```plaintext
CNT% ← 0% ! 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.

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 → described in Section 9.1.4 to indicate that the statement is spread over several physical lines.

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

```plaintext
CNT% ← 0% ! 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.

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 → described in Section 9.1.4 to indicate that the statement is spread over several physical lines.

Not all compilers adopt this strategy. Some will execute the blocks associated with every true condition in the structure.
I\% ← 0\%
LOOP
  I\% ← I\% + 1\%
  Y(I\%) ← \sin(X(I\%))
EXIT_LOOP WHEN I\%=N\%
END_LOOP

(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 (condition) DO (block of statements)

and the block is not executed at all if (condition) is false when the loop is entered. In the second
version, the test is at the end of the loop, e.g.,

REPEAT (block of statements) UNTIL (condition)

and the block will be executed at least once, whether (condition) 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 (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 . . . , 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 (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.

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.

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 IMIN% equals IMAX% and not at all if IMIN% exceeds 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

EXECUTE ⟨procedure name⟩
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@:\textsuperscript{12}

   ```
   TEMP@ ← A@
   A@ ← B@
   B@ ← TEMP@
   ! Save value in A@
   ! Copy B@ to A@
   ! Copy original A@ from TEMP@
   ```

   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]:

   ```
   SENTINEL@ ← ⟨agreed-upon special value⟩
   [COUNT% ← 0%]
   LOOP
     READ ITEM@
     EXIT_LOOP WHEN ITEM@ = SENTINEL@
     ⟨Statements processing ITEM@⟩
     [COUNT% ← COUNT% + 1%]
   END_LOOP
   ```

   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:

   ```
   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
   ```

   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:

\textsuperscript{12}Note again the use of the character \! to flag comments. See footnote 9.
9.2. SAMPLE SHORT ALGORITHMS

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,13 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

13For example, 5/2 = 2 in integer arithmetic; 5.0/2.0 = 2.5 in floating point arithmetic.
you improve the procedure on that score? Should you bother? (2) What error would occur in
the output if the EXIT_LOOP WHEN statement were placed just before the END_LOOP statement?

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_SOUUGHT@ ← (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_SOUUGHT@ 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

\[
\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = 0
\]  
(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, ... 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 $0^\circ$ and the fourth edge is maintained at $100^\circ$ 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)

$$
\frac{\partial^2 U}{\partial x^2} \approx \frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{\Delta x^2} \tag{9.2}
$$

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

$$
U_{i,j} \approx \frac{1}{4} \left( U_{i+1,j} + U_{i-1,j} + U_{i,j+1} + U_{i,j-1} \right) \tag{9.3}
$$

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.} Suppose 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

\begin{verbatim}
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
\end{verbatim}

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

\begin{verbatim}
DIMENSION U(15%,15%)
FOR I% ← 1% THRU 15% DO
    FOR J% ← 1% THRU 15% DO U(I%,J%) ← 0.0
    FOR J% ← 1% THRU 15% DO U(15%,J%) ← 100.0
FOR ITCNT% ← 1% THRU 30% DO Conduct one iteration
    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%)
\end{verbatim}

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>( n ) (number of arrays)</td>
</tr>
<tr>
<td>07:</td>
<td>2</td>
<td>( n_x )</td>
</tr>
<tr>
<td>08:</td>
<td>3</td>
<td>( n_y )</td>
</tr>
<tr>
<td>09:</td>
<td>1</td>
<td>( n_z )</td>
</tr>
<tr>
<td>10:</td>
<td>0.33</td>
<td>( A_{111} )</td>
</tr>
<tr>
<td>11:</td>
<td>0.28</td>
<td>( A_{211} )</td>
</tr>
<tr>
<td>12:</td>
<td>0.50</td>
<td>( A_{121} )</td>
</tr>
<tr>
<td>13:</td>
<td>0.42</td>
<td>( A_{221} )</td>
</tr>
<tr>
<td>14:</td>
<td>0.66</td>
<td>( A_{131} )</td>
</tr>
<tr>
<td>15:</td>
<td>0.57</td>
<td>( A_{231} )</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,\(^\text{15}\) 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
- \( n_xn_yn_z \) 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 \( n_x - 1 \); in other languages, the values will range from 1 to \( n_x \). (The subscript \( * \) 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

```
DIMENSION ARRAY(n_x, n_y)
NARR% ←− 1%
NX% ←− n_x
NY% ←− n_y
NZ% ←− 1%
DX ←− (x_{\text{max}} - x_{\text{min}})/(n_x - 1)
DY ←− (y_{\text{max}} - y_{\text{min}})/(n_y - 1)
```

Then, we evaluate the function $f(x, y)$ with the double loop expressed in the code

```
FOR J% ←− 1% THRU NY% DO
  BEGIN_BLOCK
  YF ←− y_{\text{min}} + (J%-1%)*DY
  FOR I% ←− 1% THRU NX% DO
    BEGIN_BLOCK
    XF ←− x_{\text{min}} + (I%-1%)*DX
    ARRAY(I%, J%) ←− f(XF, YF)
    END_BLOCK
  END_BLOCK
END_BLOCK
```

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\

```
ATTACH FILE (filename) ON CHANNEL 1 FOR WRITING
WRITE TO CHANNEL 1, "Line of explanation (Title, Author, Date?)"
WRITE TO CHANNEL 1, "Line of explanation"
WRITE TO CHANNEL 1, "Line of explanation"
WRITE TO CHANNEL 1, "***"
WRITE TO CHANNEL 1, "***"
WRITE TO CHANNEL 1, NARR%
WRITE TO CHANNEL 1, NX%
WRITE TO CHANNEL 1, NY%
WRITE TO CHANNEL 1, NZ%
FOR J% ←− 1% TO NY% DO FOR I% ←− 1% TO NX% DO WRITE TO CHANNEL 1, ARRAY(I%, J%)
CLOSE FILE ON CHANNEL 1
```

Here, the first statement identifies the file and assigns to it a number—any number will do\(^{16}\)—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.
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

\[
\text{ATTACH FILE } \langle \text{filename} \rangle \text{ ON CHANNEL 1 FOR READING} \\
\text{READ FROM CHANNEL 1, Comma-separated list of variables for storage of values} \\
\vdots \\
\text{CLOSE FILE ON CHANNEL 1}
\]

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.

### 9.9 Solving Laplace’s Equation with MATLAB

The algorithm described in Section 9.3.1 is readily implemented in MATLAB with the statements

\[
\begin{align*}
\text{>> u = zeros(15,15); u(:,15)=100.0;} \\
\text{>> for itcnt = 1:30} \\
\text{\quad for i = 2:14} \\
\text{\quad\quad for j = 2:14} \\
\text{\quad\quad\quad u(i,j)=0.25*( u(i+1,j)+u(i-1,j)+u(i,j+1)+u(i,j-1) );} \\
\text{\quad\end{end} \\
\text{\end{end}}
\end{align*}
\]

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 \texttt{for} loops and occupies seven physical lines) executes 30 iterations (the outermost \texttt{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.\(^{18}\)

9.12.4 ... with MATLAB

\(^{18}\)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 MATLAB, 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.F.

In MATLAB, creation of an internal array containing the values of the function of Eq. (9.4) is
quickly accomplished with the MATLAB procedure \texttt{meshgrid}. As described in Section 3.13.1, this
procedure generates two 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, we would create the desired array of values representing the function \(I(x, y)/I_0\)
with the statements\(^{19}\)

\[
\begin{align*}
\text{>> inc} & = 6.0*\text{pi}/49.0; \\
\text{>> [x,y]} & = \text{meshgrid}(-3.0*\text{pi}:\text{inc}:3.0*\text{pi}, -3.0*\text{pi}:\text{inc}:3.0*\text{pi}); \\
\text{>> I} & = ( (\sin(x)./x).^2 ) .* ( (\sin(y)./y).^2 );
\end{align*}
\]

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 MATLAB to write the array \(I\), which must already exist in the MATLAB
workspace, to a file. Before accomplishing its main objective, \texttt{lugen_list} requests input of

\[
\text{Note: All MATLAB programs (*.m) and all MATLAB-created data files (*.matlab.dat) in this}
\text{chapter can be copied from the directory $\text{HEAD}/matlab$, where (as defined in the Local Guide)
$\text{HEAD}$ must be replaced by the appropriate path for your site.}
\]

\[
\text{In MATLAB, creation of an internal array containing the values of the function of Eq. (9.4) is}
\text{quickly accomplished with the MATLAB procedure \texttt{meshgrid}. As described in Section 3.13.1, this}
\text{procedure generates two two-dimensional arrays, one containing the \(x\) coordinates of a grid of points}
\text{uniformly spaced in a region of the \(xy\) plane and the other containing the \(y\) coordinates of points}
\text{in that grid. Thus, we would create the desired array of values representing the function \(I(x, y)/I_0\)
with the statements}^{19}
\]

\[
\begin{align*}
\text{>> inc} & = 6.0*\text{pi}/49.0; \\
\text{>> [x,y]} & = \text{meshgrid}(-3.0*\text{pi}:\text{inc}:3.0*\text{pi}, -3.0*\text{pi}:\text{inc}:3.0*\text{pi}); \\
\text{>> I} & = ( (\sin(x)./x).^2 ) .* ( (\sin(y)./y).^2 );
\end{align*}
\]

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

\[
\text{>> lugen_list( I )}
\]
• 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

\[
\text{>> lugen_list( I )}
\]

All entries are strings and must be enclosed in single quotation marks.

File name [untitled.dat]: 'irrad_matlab.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_matlab.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

\[
p(x, y, z) = \frac{1}{2\pi(27)^3} \rho^4 e^{-2\rho/3} \left( \frac{3z^2}{\rho^2} - 1 \right)^2 = \frac{1}{2\pi(27)^3} e^{-2\rho/3} (9z^4 - 6z^2\rho^2 + \rho^4) \quad (9.5)
\]

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

\[
\rho = \sqrt{x^2 + y^2 + z^2} \quad (9.6)
\]

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 MATLAB procedure for reading files written by lugen_list.m is described in Section 9.16.4.
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.

### 9.13.4 ... with MATLAB

The procedure for using MATLAB to generate three-dimensional scalar arrays is similar to the method described in Section 9.12.4. The array representing the probability density in Eq. (9.5) is created using the MATLAB statements

\[
\begin{align*}
>> & \text{inc} = 20.0/29.0; \\
>> & [x,y,z] = \text{meshgrid}( -10.0: \text{inc}: 10.0, -10.0: \text{inc}: 10.0, -10.0: \text{inc}: 10.0); \\
>> & \text{pfactor} = 1.0/(2.0*\pi*27.0*27.0*27.0); \\
>> & zs = z.*z; \\
>> & \text{rhos} = x.^2 + y.^2 + zs; \\
>> & \text{rho} = \sqrt{\text{rhos}}; \\
>> & P = \text{pfactor} \times \exp(-2*\text{rho}/3) \times (9*zs.^2 - 6*zs.*\text{rhos} + \text{rhos}.^2);
\end{align*}
\]

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

\[
\begin{align*}
>> & \text{lugen\_list}(P)
\end{align*}
\]

we write the data into a file in the desired format.\(^{22}\) After the filename—we here choose \textit{pdens\_matlab.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,\(^{22}\)

\(^{22}\)The procedure \textit{lugen\_list} invoked here is identical for both two- and three-dimensional scalar and vector arrays and is described in Section 9.12.4.

\(^{23}\)A Lawrence-written MATLAB procedure for reading files written by \textit{lugen\_list\_m} is described in Section 9.16.4.
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\(^{24}\)

\[
\mathbf{E}(\mathbf{r}, t) = A \sin \left( \frac{n\pi y}{b} \right) e^{i(\kappa_x x - \omega t)} \hat{k} \tag{9.7}
\]

\[
\mathbf{H}(\mathbf{r}, t) = \frac{\nabla \times \mathbf{E}}{i\omega \mu_0} = A \left( \frac{n\pi}{b} \cos \frac{n\pi y}{b} \hat{i} - i\kappa_x \sin \frac{n\pi y}{b} \hat{j} \right) e^{i(\kappa_x x - \omega t)} \tag{9.8}
\]

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

\[
\mathbf{E}(\mathbf{r}) = A \cos \kappa_x x \sin \frac{n\pi y}{b} \hat{k}
\]

\[
\mathbf{H}(\mathbf{r}) = \frac{An\pi}{\omega \mu_0 b} \sin \kappa_x x \cos \frac{n\pi y}{b} \hat{i} - \frac{A\kappa_x}{\omega \mu_0} \cos \kappa_x x \sin \frac{n\pi y}{b} \hat{j} \tag{9.10}
\]

To simplify these equations further, we divide $\mathbf{E}$ by $A$ and $\mathbf{H}$ by $An\pi/\omega \mu_0 b$, 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

\[
\overline{\mathbf{E}}(\mathbf{r}) = \frac{\mathbf{E}(\mathbf{r})}{A} = \cos \kappa_x b \overline{x} \sin n\pi \overline{y} \hat{k} \tag{9.11}
\]

\[
\overline{\mathbf{H}}(\mathbf{r}) = \frac{\mathbf{H}(\mathbf{r})}{An\pi/\omega \mu_0 b} = \sin \kappa_x b \overline{x} \cos n\pi \overline{y} \hat{i} - \frac{\kappa_x b}{n\pi} \cos \kappa_x b \overline{x} \sin n\pi \overline{y} \hat{j}
\]

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

\[
\begin{align*}
\mathbf{E}_z &= \cos \kappa_x b \pi \sin n \pi y \\
\mathbf{H}_x &= \sin \kappa_x b \pi \cos n \pi y \\
\mathbf{H}_y &= -\frac{\kappa_x b}{n \pi} \cos \kappa_x b \pi \sin n \pi y
\end{align*}
\]  

(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

\[
\begin{align*}
\mathbf{E}_z &= \cos 2 \pi x \sin 2 \pi y \\
\mathbf{H}_x &= \sin 2 \pi x \cos 2 \pi y \\
\mathbf{H}_y &= -\cos 2 \pi x \sin 2 \pi y
\end{align*}
\]  

(9.14)

where \( \pi \) can range over any values—we choose \( 0 \leq \pi \leq 1 \)—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 \( \mathbf{H}_x(\pi, y), \mathbf{H}_y(\pi, y), \) and \( \mathbf{E}_z(\pi, y) \) over the region \( 0 \leq \pi, y \leq 1 \), dividing each axis into 30 segments of length \( 1.0 / 29.0 \), which will entail evaluating the field components at 30 values of \( \pi \) 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 \( \mathbf{E}_z \).

9.14.4 ... with MATLAB

The procedure for generating two-dimensional vector arrays is very similar to the method described in Section 9.12.4. 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 MATLAB commands

\[
\begin{align*}
&\text{>> inc = 2.0*pi/29.0;} \\
&\text{>> [x,y] = meshgrid( 0.0:inc:2.0*pi, 0.0:inc:2.0*pi );} \\
&\text{>> H(:,1,1) = sin(x).*cos(y);} \\
&\text{>> H(:,1,2) = -cos(x).*sin(y);} \\
&\text{>> H(:,1,3) = -H(:,1,2);} \\
\end{align*}
\]

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

\[
\text{>> lugen_list( H )}
\]

\[25\text{We have supposed indices starting at 1. In languages where indices start at 0, the associations would, of course, be \( \mathbf{H}(*,*,0,0) \) storing \( \mathbf{H}_x \), \( \mathbf{H}(*,*,0,1) \) storing \( \mathbf{H}_y \), and \( \mathbf{H}(*,*,0,2) \) storing \( \mathbf{E}_z \).}
\[26\text{Remember that dividing a length into 29 segments will require 30 nodes. Remember also that we are treating two-dimensional arrays as three-dimensional arrays in which the third index can assume only one value. Thus, to specify three \( 30 \times 30 \) two-dimensional arrays we must actually specify three \( 30 \times 30 \times 1 \) three-dimensional arrays.}
we write the data into a file in the desired format. After the filename—we here choose \texttt{wavegd\_matlab.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

\[
\frac{E_x(x,y,z), E_y(x,y,z), E_z(x,y,z)}{q/4\pi\epsilon_0 a^2} = \frac{[x-1,y-1,z]}{[(x-1)^2+(y-1)^2+z^2]^{3/2}} - \frac{[x+1,y-1,z]}{[(x+1)^2+(y-1)^2+z^2]^{3/2}} + \frac{[x+1,y+1,z]}{[(x+1)^2+(y+1)^2+z^2]^{3/2}} - \frac{[x-1,y+1,z]}{[(x-1)^2+(y+1)^2+z^2]^{3/2}}
\] (9.15)

To be explicit, we determine values of the field components over the region $-2 \leq x, y, z \leq 2$, 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 $30 \times 30 \times 30 \times 3$ array $E$, with $E(*,*,0,1)$ 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.

9.15.4 ... with MATLAB

The procedure for using MATLAB to generate three-dimensional vector arrays is very similar to the method described in Section 9.14.4. 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 MATLAB statements

\begin{verbatim}
>> inc = 4.0/29.0;
>> [x,y,z] = meshgrid( -2.0:inc:2.0, -2.0:inc:2.0, -2.0:inc:2.0 );
\end{verbatim}

\footnote{The procedure \texttt{ugen\_list} invoked here is identical for both two-dimensional scalar and vector arrays and is described in Section 9.12.4.}

\footnote{A Lawrence-written MATLAB procedure for reading files written by \texttt{ugen\_list.m} is described in Section 9.16.4.}

\footnote{We have supposed indices starting at 1. In languages where indices start at 0, the associations would, of course, be $E(*,*,0,0)$ storing $E_x$, $E(*,*,0,1)$ storing $E_y$, and $E(*,*,0,2)$ 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(:,:,...) evaluate the E-field components. Finally, with the statement

>> lugen_list( E )

we write the data to a file in the desired format. After the filename—we here choose quadpole_matlab.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.

9.16 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.

9.16.4 ... with MATLAB

The Lawrence-generated MATLAB function luread_list parallels the MATLAB function lugen_list first described in Section 9.12.4 and provides the facility for reading into MATLAB any file created in the format described in Section 9.3.2, whether the files were created with MATLAB or some other program. We simply execute the statement

\[ \text{Var} = \text{luread_list}( 'FileName' ) \]

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

\[ \text{>> I} = \text{luread_list}( 'irrad_matlab.dat' ); \]

will recreate the variable I as it was when the file irrad_matlab.dat was created in Section 9.12.4. 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.R.

---

30 The procedure lugen_list invoked here is identical for both two- and three-dimensional scalar and vector arrays and is described in Section 9.12.4.
31 A Lawrence-written MATLAB procedure for reading files written by lugen_list.m is described in Section 9.16.4.
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 MATLAB, see Section 3.19.

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.
Table 9.5: Algorithm for Exercise 9.5.

\[
\text{SENTINEL} \leftarrow \text{(agreed-upon special value)} \\
\text{LARGEST} \leftarrow \text{(assumed fictitious integer known to occur before any possible real integer in the list)}
\]

\[
\text{LOOP} \\
\quad \text{READ ITEM} \\
\quad \text{EXIT LOOP WHEN ITEM = SENTINEL} \\
\quad \text{IF ITEM > LARGEST} \\
\quad \quad \text{THEN LARGEST} \leftarrow \text{ITEM} \\
\quad \text{END IF} \\
\text{END LOOP} \\
\text{Report LARGEST}
\]
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 $A\%\times A\% + B\%\times B\% = C\%\times C\%$, 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 $C\%\leq A\%$ or $C\%\leq B\%$. 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

$$\frac{du}{dx}|_{x + \frac{1}{4}\Delta x} \approx \frac{u(x + \Delta x) - u(x)}{\Delta x}$$
CHAPTER 9. INTRODUCTION TO PROGRAMMING

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%
      EXITLOOP 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%
      EXITLOOP 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
\[
\frac{d^2u}{dx^2} \approx \frac{u(x + \Delta x) - 2u(x) + u(x - \Delta x)}{\Delta x^2}
\]
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
program in such a way that it asks repeatedly for Celsius temperatures until the temperature 9999 is entered, at which point the program terminates smoothly.

9.15. 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

\[
B(z) = \frac{1}{2} B_0 \left( \frac{a^2 + b^2}{a^2 + (z + b)^2} \right)^{3/2} \left[ \frac{1}{a^2 + (z + b)^2} + \frac{1}{a^2 + (z - b)^2} \right]^{3/2}
\]

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 as seen as a two-dimensional scalar function of \( z/a \) and \( b/a \). Suggestion: Write values of \( B/B_0 \) into the file.

9.16. 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.

9.17. 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

\[
p_{3,1,0}(x, y, z) = \frac{8}{(27)^2 \pi} \rho^2 \left( 1 - \frac{\rho}{\rho_0} \right)^2 e^{-2\rho/3} \cos^2 \theta
\]

\[
p_{3,1,1}(x, y, z) = \frac{4}{(27)^2 \pi} \rho^2 \left( 1 - \frac{\rho}{\rho_0} \right)^2 e^{-2\rho/3} (1 - \cos^2 \theta)
\]

\[
p_{3,2,1}(x, y, z) = \frac{3}{(27)^3 \pi} \rho^4 e^{-2\rho/3} \cos^2 \theta (1 - \cos^2 \theta)
\]

\[
p_{3,2,2}(x, y, z) = \frac{3}{4(27)^3 \pi} \rho^4 e^{-2\rho/3} (1 - \cos^2 \theta)^2
\]

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 \( \cos \theta = z/\rho \). 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

\[
p(x, y, z, t) = A \sin \frac{l \pi x}{a} \sin \frac{m \pi y}{a} \sin \frac{n \pi z}{a} \cos \omega m n t
\]

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

\[
\mathbf{E}(x, y) = \frac{q}{4 \pi \varepsilon_0} \left[ \frac{x \hat{i} + (y - a) \hat{j}}{\left[ x^2 + (y - a)^2 \right]^{3/2}} - \frac{x \hat{i} + (y + a) \hat{j}}{\left[ x^2 + (y + a)^2 \right]^{3/2}} \right]
\]

Expressing coordinates in terms of the dimensionless variables \( \tilde{x} = x/a \) and \( \tilde{y} = y/a \) and measuring \( \mathbf{E} \) in the unit \( q/(4 \pi \varepsilon_0 a^2) \), 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

\[
\mathbf{v} = \mathbf{\omega} \times \mathbf{r} = \omega \hat{k} \times (x \hat{i} + y \hat{j} + z \hat{k}) = \omega (-y \hat{i} + x \hat{j})
\]

Choosing a unit of length \( a \) and expressing coordinates in units of \( a \) and velocities in units of \( \omega a \), 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

\[
\mathbf{E}(x, y, z) = \frac{q}{4 \pi \varepsilon_0} \left[ \frac{x \hat{i} + y \hat{j} + (z - a) \hat{k}}{\left[ x^2 + y^2 + (z - a)^2 \right]^{3/2}} - \frac{x \hat{i} + y \hat{j} + (z + a) \hat{k}}{\left[ x^2 + y^2 + (z + a)^2 \right]^{3/2}} \right]
\]

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 \( q/(4 \pi \varepsilon_0 a^2) \), 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 \(43°40' N, 73°45' W\); Grand Junction, CO, is at \(39°55' N, 108°33' W\); Los Angeles, CA, is at \(34°3' N, 118°15' W\); Appleton, WI, is at \(44°16' N, 88°25' W\); Calcutta India, is at \(22°32' N, 88°20' E\); Sydney, Australia, is at \(33°52' S, 151°12' E\); Paris, France, is at \(48°49' N, 2°29' E\); and Stockholm, Sweden, is at \(59°21' N, 18°4' E\).
function [] = lugen_list(array)
% LUGEN_LIST writes data in an array to a file
% LUGEN_LIST facilitates writing data stored in an array in
% MATLAB 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, '
All entries are strings and must be enclosed
')
fprintf(1, 'in single quotation marks.

')
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\n', title );
fprintf( id, '%s\n', comm1 );
fprintf( id, '%s\n', comm2 );
fprintf( id, '%s\n', comm3 );
fprintf( id, '%s\n', comm4 );
fprintf( id, '%d\n', narr );
fprintf( id, '%d\n', xdim );
fprintf( id, '%d\n', ydim );
fprintf( id, '%d\n', 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, '
Reading File
' )
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, ['
File ''',FileName,''' read successfully!
'] )
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 \( |F_v| = f(|\mathbf{v}|) \). In general, if the projectile has mass \( m \), Newton’s second law yields the equation of motion

\[
m \frac{d^2 \mathbf{r}}{dt^2} = -mg \hat{k} + F_v(|\mathbf{v}|) = -mg \hat{k} - f(|\mathbf{v}|) \frac{\mathbf{v}}{|\mathbf{v}|}
\]

which we are to solve subject to the general initial conditions,

\[
\mathbf{r}(0) = \mathbf{r}_0 = x_0 \hat{i} + y_0 \hat{j} + z_0 \hat{k} \quad ; \quad \mathbf{v}(0) = \mathbf{v}_0 = v_{x0} \hat{i} + v_{y0} \hat{j} + v_{z0} \hat{k}
\]  

(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

\[
m \frac{d^2 \mathbf{r}}{dt^2} = -mg \hat{k} - b \mathbf{v} = -mg \hat{k} - b \frac{d \mathbf{r}}{dt}
\]

or, in component form, to

\[
m \frac{d^2 x}{dt^2} = -b \frac{dx}{dt} \quad ; \quad m \frac{d^2 y}{dt^2} = -b \frac{dy}{dt} \quad ; \quad m \frac{d^2 z}{dt^2} = -mg - b \frac{dz}{dt}
\]

(11.3)

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

\[
m \frac{d^2 \mathbf{r}}{dt^2} = -mg \hat{k} - b |\mathbf{v}| = -mg \hat{k} - b \frac{d \mathbf{r}}{dt} \sqrt{\left( \frac{dx}{dt} \right)^2 + \left( \frac{dy}{dt} \right)^2 + \left( \frac{dz}{dt} \right)^2}
\]

or, in component form, to

\[
m \begin{bmatrix} \frac{d^2 x}{dt^2}, \frac{d^2 y}{dt^2}, \frac{d^2 z}{dt^2} \end{bmatrix} = \begin{bmatrix} 0, 0, -mg \end{bmatrix} - b \begin{bmatrix} \frac{dx}{dt}, \frac{dy}{dt}, \frac{dz}{dt} \end{bmatrix} \sqrt{\left( \frac{dx}{dt} \right)^2 + \left( \frac{dy}{dt} \right)^2 + \left( \frac{dz}{dt} \right)^2}
\]

(11.4)

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

\[
m \frac{d^2 z}{dt^2} = -mg - b \frac{dz}{dt} \left| \frac{dz}{dt} \right|
\]

(11.5)

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 \( a \)—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 \( \vec{x} = x/a, \vec{y} = y/a, \) and \( \vec{z} = z/a \) (i.e., \( \vec{r} = r/a \)), we find that Eqs. (11.3) and (11.5) become

\[
\frac{d^2 \vec{r}}{dt^2} = -\frac{g}{a} \vec{k} - \frac{b}{m} \frac{d\vec{r}}{dt}
\]

and

\[
\frac{d^2 \vec{r}}{dt^2} = -\frac{g}{a} \vec{k} - \frac{ba}{m} \frac{d\vec{r}}{dt} \sqrt{\left(\frac{d\vec{x}}{dt}\right)^2 + \left(\frac{d\vec{y}}{dt}\right)^2 + \left(\frac{d\vec{z}}{dt}\right)^2}
\]

respectively. Next, we recognize that \( \sqrt{g/a} \) is dimensionally a frequency.\(^2\) Hence, the variable \( t = t/\sqrt{g/a} \) provides a suitable rescaling of the independent variable. In terms of \( \vec{r} \), Eqs. (11.8) and (11.9) become

\[
\frac{d^2 \vec{r}}{d\vec{t}^2} = -\dot{k} - \frac{b}{m} \sqrt{\frac{a}{g}} \frac{d\vec{r}}{d\vec{t}}
\]

and

\[
\frac{d^2 \vec{r}}{d\vec{t}^2} = -\dot{k} - \frac{ba}{m} \sqrt{\frac{a}{g}} \frac{d\vec{r}}{d\vec{t}} \sqrt{\left(\frac{d\vec{x}}{d\vec{t}}\right)^2 + \left(\frac{d\vec{y}}{d\vec{t}}\right)^2 + \left(\frac{d\vec{z}}{d\vec{t}}\right)^2}
\]

Finally, we introduce the symbol \( \vec{\beta} \) 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

\[
\frac{d^2 \vec{r}}{d\vec{t}^2} = -\dot{k} - \vec{\beta} \frac{d\vec{r}}{d\vec{t}}
\]

and

\[
\frac{d^2 \vec{r}}{d\vec{t}^2} = -\dot{k} - \vec{\beta} \sqrt{\frac{a}{g}} \frac{d\vec{r}}{d\vec{t}} \sqrt{\left(\frac{d\vec{x}}{d\vec{t}}\right)^2 + \left(\frac{d\vec{y}}{d\vec{t}}\right)^2 + \left(\frac{d\vec{z}}{d\vec{t}}\right)^2}
\]

These equations are, of course, to be solved subject to the dimensionless initial conditions

\[
\vec{r}(0) = \frac{r_0}{a} \quad ; \quad \vec{v}(0) = \frac{d\vec{r}}{dt}(0) = \frac{d(\vec{r}/a)}{dt}(0) = \frac{v_0}{\sqrt{g/a}}
\]

from which we conclude that dimensionless velocities are measured in units of \( \sqrt{g/a} \).

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

\[
m \frac{dv_x}{dt} = -bv_x ; \quad m \frac{dv_y}{dt} = -bv_y ; \quad m \frac{dv_z}{dt} = -mg - bv_z
\]

the three members of Eq. (11.6) become

\[
m \left[ \frac{dv_x}{dt}, \frac{dv_y}{dt}, \frac{dv_z}{dt} \right] = [0, 0, -mg] - b[v_x, v_y, v_z] \sqrt{v_x^2 + v_y^2 + v_z^2}
\]

and Eq. (11.7) becomes

\[
m \frac{dv_z}{dt} = -mg - bv_z \ |v_z|
\]

Once these first-order equations have been solved, the components of the position vector can then be found by solving the differential equations

\[
\frac{dx}{dt} = v_x ; \quad \frac{dy}{dt} = v_y ; \quad \frac{dz}{dt} = v_z
\]

\(^2\) 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

Figure 11.2: A three-nucleus radioactive decay.

\[
\begin{array}{c}
\text{A} & \xrightarrow{k_A} & \text{B} & \xrightarrow{k_B} & \text{C}
\end{array}
\]

By the time we have reached this point, however, we will have explicit knowledge of \(v_x, v_y, \) and \(v_z\) as functions of \(t\), so solving Eq. (11.18) is equivalent to straightforward evaluation of an integral. (See Chapter 13.) Recasting the equations of this paragraph in dimensionless form is left as an exercise.

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

\[
\begin{align*}
\frac{dA}{dt} &= -k_A A \\
\frac{dB}{dt} &= k_A A - k_B B \\
\frac{dC}{dt} &= k_B B
\end{align*}
\]

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:\footnote{We shall later see that conservation laws can sometimes prove valuable in assessing the accuracy of solutions.} adding the three equations yields that

\[
\frac{d}{dt}(A + B + C) = 0 \implies A + B + C = \text{constant}
\]

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 \(\bar{A} = A/A_0, \) \(\bar{B} = B/A_0, \) and \(\bar{C} = C/A_0.\) Next, dividing the equations by \(k_A A_0\) and introducing the dimensionless quantities \(\bar{t} = k_A t\) and \(\bar{k} = k_B/k_A,\) we conclude that

\[
\begin{align*}
\frac{d\bar{A}}{dt} &= -\bar{A} \\
\frac{d\bar{B}}{dt} &= \bar{A} - \bar{k} \bar{B} \\
\frac{d\bar{C}}{dt} &= \bar{k} \bar{B}
\end{align*}
\]

with the initial conditions \(\bar{A}(0) = 1, \bar{B}(0) = 0, \) and \(\bar{C}(0) = 0, \) where we now regard the dependent variables as functions of the dimensionless time \(\bar{t}.\) In short, we discover that this problem possesses only one “real” parameter \(\bar{k}.\) 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.
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

\[
\frac{dN}{dt} = kN \left(1 - \frac{N}{N_c}\right)
\]

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.

The dimensionless version of this equation is quickly found. We choose \(N_c\) as the reference population, introduce the dimensionless population \(\bar{N} = N/N_c\) and the dimensionless time \(\bar{t} = kt\), and find that the equation and initial condition reduce to

\[
\frac{d\bar{N}}{d\bar{t}} = \bar{N}(1 - \bar{N}) ; \quad \bar{N}(0) = \frac{N_0}{N_c}
\]

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

\[
m \frac{d^2x}{dt^2} = -kx - b \frac{dx}{dt} + F(t)
\]

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

\[
x(0) = x_0 ; \quad v(0) = v_0
\]

before the problem is fully stated.

---

6Logistic 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 $x = x/\ell$, introduce a unit of time, say $\tau$, rescale the physical time with the expressions $t = t/\tau$, and thereby transform Eq. (11.24) to

$$\frac{d^2x}{dt^2} = -\frac{\tau^2 k}{m} x - \frac{\tau b}{m} \frac{dx}{dt} + \frac{\tau^2 F(\tau t)}{m\ell}$$

(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

$$\frac{d^2x}{dt^2} = -x - \beta \frac{dx}{dt} + F(t)$$

(11.27)

where, in the final form, we have set $\beta = b/\sqrt{mk}$ and $F(t) = F(\tau t)/k\ell$. 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 $F$ 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

$$\bar{x}(0) = \frac{x_0}{\ell} ; \quad \bar{v}(0) = \frac{d\bar{x}}{dt}(0) = \frac{d(x/\ell)}{d(t/\tau)} = \frac{v_0}{\ell/\tau}$$

(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

$$\frac{dx}{dt} = v ; \quad m \frac{dv}{dt} = -kx - bv + F(t)$$

(11.29)

in the original dimensional form, or

$$\frac{d\bar{x}}{dt} = \bar{v} ; \quad \frac{d\bar{v}}{dt} = -\bar{x} - \beta \bar{v} + F(\bar{t})$$

(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

$$V(t) - L \frac{di}{dt} - \frac{q}{C} - iR = 0 \implies L \frac{di}{dt} + iR + \frac{q}{C} = V(t) \quad (11.31)$$

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

$$i = \frac{dq}{dt} \quad (11.32)$$

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

$$q(0) = q_0 ; \quad \frac{dq}{dt}(0) = i(0) = i_0 \quad (11.33)$$

Alternatively, we could substitute Eq. (11.32) into Eq. (11.31) to find the equivalent, single, second-order, linear, constant coefficient, inhomogeneous equation

$$L \frac{d^2q}{dt^2} + R \frac{dq}{dt} + \frac{q}{C} = V(t) \quad (11.34)$$

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.\footnote{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 \textit{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.}

11.1.6 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.
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

$$m \frac{d^2 x_1}{dt^2} = -kx_1 + k'(x_2 - x_1) \quad ; \quad m \frac{d^2 x_2}{dt^2} = -kx_2 - k'(x_2 - x_1)$$

(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 $x_i/a = \bar{x}_i$, introducing a dimensionless time variable $\bar{t} = \omega t$, where $\omega = \sqrt{k/m}$, and setting $\bar{k} = k'/k$. The equations then become

$$\frac{d^2 \bar{x}_1}{d\bar{t}^2} = -\bar{x}_1 + \bar{k} (\bar{x}_2 - \bar{x}_1) \quad ; \quad \frac{d^2 \bar{x}_2}{d\bar{t}^2} = -\bar{x}_2 - \bar{k} (\bar{x}_2 - \bar{x}_1)$$

(11.36)

Initial conditions such as

$$\bar{x}_1(0) = \bar{x}_{10} \quad ; \quad \frac{d\bar{x}_1}{d\bar{t}}(0) = \bar{v}_{10} \quad ; \quad \bar{x}_2(0) = \bar{x}_{20} \quad ; \quad \frac{d\bar{x}_2}{d\bar{t}}(0) = \bar{v}_{20}$$

(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 $\bar{k}$.

### 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.
CHAPTER 11. SOLVING ORDINARY DIFFERENTIAL EQUATIONS

the differential equation
\[ m \frac{d^2 \mathbf{r}}{dt^2} = \mathbf{F} = f(r) \mathbf{\hat{r}} \] (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 \( r^2 = x^2 + y^2 \), \( \mathbf{r} = x \mathbf{i} + y \mathbf{j} \), and
\[ \mathbf{\hat{r}} = \frac{\mathbf{r}}{r} = \frac{x \mathbf{i} + y \mathbf{j}}{\sqrt{x^2 + y^2}} \] (11.39)

the Cartesian components of Eq. (11.38) are
\[ m \frac{d^2 x}{dt^2} = f \left( \sqrt{x^2 + y^2} \right) \frac{x}{\sqrt{x^2 + y^2}} ; \quad m \frac{d^2 y}{dt^2} = f \left( \sqrt{x^2 + y^2} \right) \frac{y}{\sqrt{x^2 + y^2}} \] (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,\(^9\) 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
\[ \frac{d^2 x}{dt^2} = -GM \frac{x}{(x^2 + y^2)^{3/2}} ; \quad \frac{d^2 y}{dt^2} = -GM \frac{y}{(x^2 + y^2)^{3/2}} \] (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
\[ \overline{x} = \frac{x}{\ell} ; \quad \overline{y} = \frac{y}{\ell} ; \quad \overline{r} = \frac{r}{\ell} \] (11.44)

With this change, Eq. (11.43) becomes
\[ \frac{d^2 \overline{x}}{d\overline{t}^2} = -\frac{GM}{\ell^3} \frac{\overline{x}}{(\overline{x}^2 + \overline{y}^2)^{3/2}} ; \quad \frac{d^2 \overline{y}}{d\overline{t}^2} = -\frac{GM}{\ell^3} \frac{\overline{y}}{(\overline{x}^2 + \overline{y}^2)^{3/2}} \] (11.45)

Finally, we introduce the dimensionless time variable \( \overline{t} = t \sqrt{GM/\ell^3} \) to find that
\[ \frac{d^2 \overline{x}}{d\overline{t}^2} = -\frac{\overline{x}}{(\overline{x}^2 + \overline{y}^2)^{3/2}} ; \quad \frac{d^2 \overline{y}}{d\overline{t}^2} = -\frac{\overline{y}}{(\overline{x}^2 + \overline{y}^2)^{3/2}} \] (11.46)

Interestingly, in Cartesian coordinates, all parameters have disappeared from the equations.

\(^9\)Remember, 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
\[ x(0) = \frac{x_0}{\ell}; \quad y(0) = \frac{y_0}{\ell} \] (11.47)

To translate the velocities, we argue that
\[ \frac{d}{dt}(0) = \frac{d(x/\ell)}{d(t/\sqrt{GM/\ell^3})}(0) = \frac{v_{x0}}{\sqrt{GM/\ell}}; \quad \frac{d}{dt}(0) = \frac{d(y/\ell)}{d(t/\sqrt{GM/\ell^3})}(0) = \frac{v_{y0}}{\sqrt{GM/\ell}} \] (11.48)

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
\[ \frac{mv_{\text{circ}}^2}{r_{\text{circ}}} = \frac{GMm}{r_{\text{circ}}^2} \implies v_{\text{circ}}^2 = \frac{GM}{r_{\text{circ}}} \] (11.49)

Translated into dimensionless form, this special relationship becomes
\[ \left( \frac{\sqrt{GM/\ell}}{v_{\text{circ}}} \right)^2 = \frac{GM}{\ell v_{\text{circ}}} \implies v_{\text{circ}}^2 = \frac{1}{r_{\text{circ}}} \] (11.50)

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
\[ E = \frac{1}{2}m(v_x^2 + v_y^2) - \frac{GMm}{\sqrt{x^2 + y^2}} = \text{constant} \] (11.51)

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
\[ L = m(xv_y - yv_x) = \text{constant} \] (11.52)

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
\[ \frac{1}{2}(v_x^2 + v_y^2) - \frac{1}{\sqrt{\bar{x}^2 + \bar{y}^2}} = \text{constant} ; \quad \bar{x}v_y - \bar{y}v_x = \text{constant} \] (11.53)

### 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
\[ \frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \] (11.54)
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

\[
\frac{d^2 f}{dx^2} + k^2 f = 0
\]

where \( k^2 = \omega^2 / c^2 \). 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 \( k^2 = 2mE/\hbar^2 \) (or \( E = \hbar^2 k^2 / 2m \)), 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,

\[
\frac{d^2 \psi}{d\xi^2} + (2\epsilon - \xi^2) \psi = 0
\]

where \( \omega = \sqrt{k/m} \), \( \epsilon = E/\hbar \omega \) and \( \xi = x/\sqrt{\hbar/m\omega} \). In contrast to the infinitely deep well, the domain for the quantum oscillator extends over the interval \( -\infty < x < +\infty \), 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 $x$\textsuperscript{10}. Thus, the solutions to the equation can be divided into two sets, one of which contains functions that are even in $x$ (even parity) and the other of which contains functions that are odd in $x$ (odd parity). 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.\textsuperscript{11} 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

\[
\frac{d^2 \psi}{dx^2} + (2\epsilon - x^2) \psi = 0 ; \quad \psi(0) = 1.0 ; \quad \frac{d\psi}{dx}(0) = 0 ; \quad \lim_{x \to \infty} \psi(x) = 0
\] (11.62)

and yielding even solutions and the other defined by the equation and initial values

\[
\frac{d^2 \psi}{dx^2} + (2\epsilon - x^2) \psi = 0 ; \quad \psi(0) = 0 ; \quad \frac{d\psi}{dx}(0) = 1 ; \quad \lim_{x \to \infty} \psi(x) = 0
\] (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 $\tau = +\infty$ 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

\[
m[s^2 \tilde{z}(s) - sz_0 - vz_0] = -\frac{mg}{s} - b[s \tilde{z}(s) - z_0]
\] (11.65)

Then we solve this algebraic equation\textsuperscript{12} for $\tilde{z}(s)$, finding that

\[
\tilde{z}(s) = \frac{z_0}{s} + \frac{vz_0 - g}{s^2(s + b/m)}
\] (11.66)

Finally, we invert this Laplace transform to find the solution itself. Unfortunately, inverting Laplace transforms is rarely easy.\textsuperscript{13} We can, of course, read a table of transforms such as Table 1.2 backwards.

\textsuperscript{10}It involves only second derivatives and the potential energy involves only $x^2$.

\textsuperscript{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.

\textsuperscript{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.

\textsuperscript{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

\[
\hat{z}(s) = \frac{z_0}{s} - \frac{mg/b}{s^2} + \frac{m}{b} \left( v_{z0} + \frac{mg}{b} \right) \left( \frac{1}{s} - \frac{1}{s + b/m} \right)
\]  

(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

\[
z(t) = z_0 - \frac{mg}{b} t + \frac{m}{b} \left( v_{z0} + \frac{mg}{b} \right) \left( 1 - e^{-bt/m} \right)
\]  

(11.68)

Note that, in this approach, the initial conditions imposed on the solution are incorporated \textit{ab initio}, not imposed \textit{after} a general solution with arbitrary, unknown constants has been obtained.

### 11.3 Solving ODEs Symbolically with MAXIMA

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 of the commands available in MAXIMA for performing these operations were discussed briefly in Section 6.7.8. The full spectrum of commonly used commands is conveniently divided into two categories. Those commands that are useful in addressing a \textit{single} equation of whatever order include

\[
\text{depends( dvar1, ivar1, dvar2, ivar2, ... );}
\]

This command, in which each argument may be a single variable or a list, informs MAXIMA that certain quantities depend on other quantities—a specification necessary for the routines that solve single ODEs. The syntax is illustrated more explicitly in the examples below. The command \text{depends} must be invoked \textit{before} \text{ode2}.

\[
\text{ode2( eqn, depvar, indvar );}
\]

which (despite the 2 in its name) requests the solution of the first- or second-order ordinary differential equation \text{eqn} for the dependent variable \text{depvar} as a function of the independent variable \text{indvar}. \text{ode2} actually invokes a succession of increasingly more complicated methods until it finds one that works. Available system variables, which provide information about the process once a solution has been obtained, include

- \text{method}, which is set to the method ultimately used.
- \text{intfactor}, which is set to any integrating factor used and will have the value \text{false} if no integrating factor is used.
- \text{yp}, which is set to the particular solution used if variation of parameters is the method and will not be set at all if that method is not used.

Each of these variables can be displayed to reveal its value. Further, solutions returned by \text{ode2} use the variable \%c for the integration constant in the solution to a first-order equation and the variables \%k1 and \%k2 for the integration constants in the solution to a second-order equation.
11.3. SOLVING ODES SYMBOLICALLY WITH MAXIMA

ic1(sol, xval, yval)

which imposes an initial condition on the solution sol obtained by ode2 for a first-order equation, i.e., determines a specific (symbolic) value for the constant %c. xval is a statement of the form x = x0 specifying a value of the independent variable, and yval is a statement of the form y = y0 specifying the value of the dependent variable at the given value of the independent variable.

ic2(sol, xval, yval, ydrvval)

which imposes initial conditions on the solution sol obtained by ode2 for a second-order equation, i.e., determines specific (symbolic) values for the constants %k1 and %k2. xval is a statement of the form x = x0 specifying a value of the independent variable, yval and ydrvval are statements of the form y = y0 and diff(y,x) = ydrv0 specifying the values of the dependent variable and its first derivative at the given value of the independent variable.

bc2(sol, xval1, yval1, xval2, yval2)

which imposes boundary conditions on the solution sol obtained by ode for a second-order equation, i.e., determines specific (symbolic) values for the constants %k1 and %k2. Here, xval1 and xval2 are statements of the form x = x1 and x = x2 specifying two different values of the independent variable; and yval1 and yval2 are statements of the form y = y1 and y = y2 specifying the values of the dependent variable at the two given values of the independent variable.

MAXIMA does not appear to have the capacity to solve ODEs when one or both boundary conditions impose constraints on the derivatives of the functions.

To address systems of linear equations, MAXIMA uses the Laplace transform (see Sections 1.5.2, 6.7.7 and 11.2) and, thus, incorporates the initial conditions as the solutions are found. Behind the scenes, MAXIMA invokes the already-described\footnote{See Section 6.7.7.} commands laplace and ilt (and, occasionally, we may have to do the same). To support a more direct attack, MAXIMA provides two commands that are specifically built for solving systems of linear, ordinary differential equations:

\texttt{desolve( [list of equations], [list of dependent variables]);}

which requests the solution of the system of linear equations provided in the first list for the variables provided in the second list.

\texttt{atvalue( function, indval, functionval );}

which provides the value functionval to be imposed by desolve on the function function at the value of the independent variable specified by indval. Here, function must include the independent variable as an \texttt{explicit} argument and may be either the dependent variable itself or some derivative of that dependent variable, indval is a statement of the form x = 0 specifying a value of the independent variable, and functionval is an expression stipulating the value to be assumed by function at the specified point. The syntax of this command is more clearly illustrated in the specific examples below. The command atvalue must be invoked \texttt{before} desolve.

The command \texttt{depends} plays no role in dealing with systems of equations; all dependences must be \texttt{explicitly} displayed, both in \texttt{desolve} and in \texttt{atvalue}.\footnote{See Section 6.7.7.}
The MAXIMA distribution also includes the command **contrib_ode**, which can be loaded into the MAXIMA workspace with the command `load('contrib_ode');`. The additional features, which include a capacity to apply other methods to solve linear and non-linear first-order ODEs and linear homogeneous second-order ODEs, are described in the MAXIMA manuals. The syntax is the same as that for `ode2`. Be aware, however, that (as of May 2016) this package has not yet stabilized so it should be used with caution. Further, once it has stabilized, it may be moved to be available without invoking the `load` command.

### 11.3.1 Projectile in a Viscous Medium

For the sake of a comparison with the result obtained in Section 11.2, we begin by using MAXIMA to solve the third member of Eq. (11.4). The necessary commands are\(^{15,16}\)

\begin{align*}
\text{(\%i1)} & \text{depends( } z, t \text{ )} \\
\text{(\%i2)} & \text{eq : m*'diff(z,t,2)=-m*g-b*'diff(z,t)} \\
\text{(\%i3)} & \text{soln : ode2( eq, z, t );}
\end{align*}

Is \( b \) zero or nonzero? \( n; \)

\begin{align*}
\text{(\%o3)} & z = \frac{\%k2 e^{-bt/m}}{b^2} - \frac{b g m t - g m^2}{b^2} + \%k1 \\
\text{(\%i4) method;} \\
\text{(\%o4) variationofparameters}
\end{align*}

\begin{align*}
\text{(\%i5) ic2( \%o3, t=0, z=z0, 'diff(z,t)=vz0 );} & \text{ Impose initial conditions.} \\
\text{(\%o5)} & z = \frac{b z_0 + m v z_0}{b} - \frac{e^{-bt/m} (b m v z_0 + g m^2)}{b^2} - \frac{b g m - g m^2}{b^2}
\end{align*}

Unfortunately, this form of the solution is not particularly transparent. Cleaning up the expression can be tricky, and success is likely to depend on having an \textit{a priori} sense of the desired end result. The route to adopt will be clearer if we invoke the statement

\begin{align*}
\text{(\%i6) soln : expand( rhs(\%) );} \\
\text{(\%o6)} & z_0 - \frac{m e^{-bt/m} v z_0}{b} + \frac{m v z_0}{b} - \frac{g m^2 e^{-bt/m}}{b^2} - \frac{g m t}{b} + \frac{g m^2}{b^2}
\end{align*}

to extract the right-hand side of the solution at \%o5 and expand the result so that the individual terms will be clearer. In this form, we note that two terms have a common factor of \( g m^2/b^2 \) and two other terms have a common factor of \( m v z_0/b \). Because the two terms containing the factor \( e^{-bt/m} \) are not stored internally in a way that allows isolation of the factors of interest, a statement like

\begin{align*}
\text{collectterms(soln, gm^-2/b^-2);}
\end{align*}

will not achieve the desired rearrangement. Instead, we must (temporarily) replace \( g m^2/b^2 \) with, say, \( \alpha \) and \( m v z_0/b \) with, say, \( \beta \) before `collectterms` can be invoked successfully. Even those temporary replacements, however, must be done in a way that makes sure MAXIMA can, in fact, identify the quantities to be replaced in its internal storage. Thus, substituting \( \alpha \) as desired requires that we invoke the statements

\begin{itemize}
\item As in Chapter 6, we abbreviate the presentation of MAXIMA dialogs by making liberal use of terminating dollar signs to suppress intermediate output. You are urged to duplicate the dialog in an actual session with MAXIMA, replacing all terminating dollar signs with semi-colons.
\item See Section 6.7.8 for a discussion of the use of the tic in the present context.
\end{itemize}
11.3. SOLVING ODES SYMBOLICALLY WITH MAXIMA

(%i7) subst( alpha*b^2/g, m^2, %);
(%o7)

(%i8) subst( beta*b/m, vz0, %);
(%o8)

(%i9) collectterms(%, alpha, beta);
(%o9)

(%i10) collectterms( %, 1-%e^(-b*t/m) );
(%o10)

(%i11) soln : ev( %, alpha = g*m^2/b^2, beta = m*vz0/b );
(%o11)

We can now also see the common factor of \( m/b \) in the second parentheses, so an even cleaner form of the solution would be

\[
z(t) = z_0 - \frac{mg}{b} t + \frac{m}{b} \left( vz_0 + \frac{mg}{b} \right) \left( 1 - e^{-bt/m} \right) \tag{11.69}
\]

but we will not struggle to force MAXIMA to achieve this last rearrangement. This result agrees fully with the result obtained in Eq. (11.68).

Wisdom suggests that we should whenever possible actually verify any solution that MAXIMA generates. We can substitute the solution into the original equation and evaluate the derivatives explicitly with the statement

(%i12) expand( ev( eq, z=soln, diff ) );
(%o12)

The left and right sides of this result are clearly equal, though we could make that more explicit either by using the statement

(%i13) lhs(%) - rhs(%) ;
(%o13) 0

or by invoking the test

(%i14) is(%o12 ) ;
(%o14) true

where, with %o12 having the form \( a = b \), the command is simply returns true or false depending on whether the specified expression is, in fact, true or false. To complete the verification, we check satisfaction of the initial conditions with the statement

(%i15) vel : diff( soln, t )$
(%i16) expand( ev([soln, vel], t=0 ) );
(%o16) \[z_0, vz_0]
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

(%i17) expand( taylor( soln, b, 0, 1 ) );
(%o17) $z_0 + v_{z_0} t - \frac{1}{2} g t^2 - \frac{b t^2}{2m} \left( v_{z_0} - \frac{1}{2} g t \right)$

(%i18) expand( taylor( vel, b, 0, 1 ) );
(%o18) $v_{z_0} - g t - \frac{b t}{m} \left( v_{z_0} - \frac{1}{2} g t \right)$

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.

### 11.3.2 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 MAXIMA statements

(%i1) depends( N, t );
(%i2) eq : 'diff( N, t ) = k*N*(1-N/N_C);
(%i3) ode2( %, N, t ); Establish dependence of $N$ on $t$.
Enter equation.
Solve equation for $N$ as function of $t$.

(%o3) $-\frac{\log(N - N_c) - \log(N)}{k} = t + %c$

Note that MAXIMA has returned an implicit solution!

(%i4) ic1( %, t = 0, N = N0 ); Impose general initial condition.

(%o4) $-\frac{\log(N - N_c) - \log(N)}{k} = kt - \log(N_0 - N_c) + \log(N_0)$

(%i5) method; Display method used.
(%o5) separable

Strictly, of course, we want $N$ expressed as an explicit function of $t$. The solve command, however, will not by itself solve %o4 for $N$. We must guide MAXIMA with the statements

(%i6) part( k*solve( %o4, t ), 1 ); Solve for $kt$ as first step in solving for $N$.
Because solve apparently cannot solve for $kt$ directly, we invoke first solve for $t$ and then multiply by $k$. Because solve returns the solution as a one-element list, we use the part command to extract that element, facilitating later operations.

(%o6) $kt = \log(N_0 - N_c) - \log(N - N_c) - \log(N_0) + \log(N)$

---

$^{17}$Again, we have taken the liberty to rearrange the terms in MAXIMA’s output to facilitate recognition of the limits.
11.3. SOLVING ODES SYMBOLICALLY WITH MAXIMA

(\%i7) \text{logcontract}( \% );

Compress logarithms.

(\%i7) \text{kt} = \log\left(\frac{N(N_c - N_0)}{N_0(N_c - N)}\right)

(\%i8) \text{map}( \exp, \% );

Remove logarithms by using the map function to apply \exp separately to each side of \%o7.

(\%i8) e^{kt} = \frac{N(N_c - N_0)}{N_0(N_c - N)}

(\%i9) \text{part}( \text{solve}( \%, N ), 1 );

Solve (finally) for N.

(\%i9) N = \frac{N_0 N_c e^{kt}}{N_0 e^{kt} + N_c - N_0}

(\%i10) \exp( \%, t = 0 );

Verify satisfaction of initial condition.

(\%i10) N = N_0

(\%i11) \text{lhs}( \text{eq} ) - \text{rhs}( \text{eq} ) = 0$

Rearrange equation ...

(\%i12) \exp( \%, \%o9, \text{diff}, \text{radcan} );$

... and verify that solution satisfies it.

\textbf{Note:} Here, the versatility of the function \exp is again demonstrated. On line \%i12, we have told MAXIMA to evaluate \%o11 by substituting the solution from line \%o9, explicitly evaluating any occurrences of the operator \text{diff}, and finally applying the operator \text{radcan} to the result. Thus, in one statement, we have told MAXIMA to accomplish several operations.

(\%i12) 0 = 0

(\%i13) \text{limit}( \%o9, t, \text{inf});

Examine long-term solution.

Is k positive, negative, or zero? p;

(\%i13) N = N_c

When k > 0, solution approaches carrying capacity, regardless of initial population.

Is k positive, negative, or zero? n;

(\%i14) N = 0

When k < 0, population becomes extinct.

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 extract the function to be plotted and recast it with the statements

(\%i15) \text{rhs}(\%o9) / \text{NC};

Recast to measure population in units of \(N_c\).

(\%i15) \frac{N_0 e^{kt}}{N_0 e^{kt} + N_c - N_0}

(\%i16) \text{factor}( \text{subst}( \alpha \text{\* NC}, N_0, \% ) );$

Recast to measure initial population in units of \(N_c\).

(\%i16) \frac{\alpha e^{kt}}{\alpha e^{kt} - \alpha + 1}$

(\%i17) \text{soln : subst}( \tau / k, t, \% );$

Finally, recast to measure time in units of \(1/k\).

(\%i17) \frac{\alpha e^{\tau}}{\alpha e^{\tau} - \alpha + 1}$

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

(\%i18) \text{soln1: ev}( \text{soln}, \alpha = 0.25 )$

(\%i19) \text{soln2: ev}( \text{soln}, \alpha = 0.5 )$

(\%i20) \text{soln3: ev}( \text{soln}, \alpha = 1.0 )$

(\%i21) \text{plot}( \%o18, \%o19, \%o20 );$

(\%i21) \text{plot}( \%o18, \%o19, \%o20 );$

(\%i22) \text{plot}( \%o18, \%o19, \%o20 );$

(\%i23) \text{plot}( \%o18, \%o19, \%o20 );$

(\%i24) \text{plot}( \%o18, \%o19, \%o20 );$

(\%i25) \text{plot}( \%o18, \%o19, \%o20 );$

(\%i26) \text{plot}( \%o18, \%o19, \%o20 );$

(\%i27) \text{plot}( \%o18, \%o19, \%o20 );$
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, which—because $N(0)/N_c = \alpha$, is also the value of $\alpha$ for the graph.

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

```maxima
(%i24) plot2d( ev( soln, alpha=0.05 ), [tau, 0.0, 10.0], [y, 0.0, 1.0],
[legend, false], [style, [lines, 4]], [color, black],
[xlabel, "tau"], [ylabel, "N(tau) / 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.

### 11.3.3 Damped Harmonic Oscillator

A third example of the use of MAXIMA to solve a single second-order, linear, constant-coefficient, homogeneous, ordinary differential equation was presented as an example in Chapter 6. You are invited to review the discussion of the damped harmonic oscillator in Sections 6.7.8 and 11.1.4.

### 11.3.4 Chain Radioactive Decay

As a first example of the use of MAXIMA 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
11.3. SOLVING ODES SYMBOLICALLY WITH MAXIMA

Figure 11.8: Logistic growth when the population is initially much smaller than the carrying capacity. This curve is called the sigmoid curve.

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 MAXIMA statements

(%i1) eqa : 'diff( A(t), t ) = -kA*A(t)$
(%i2) eqb : 'diff( B(t), t ) = kA*A(t)-kB*B(t)$
(%i3) eqc : 'diff( C(t), t ) = kB*B(t)$
(%i4) atvalue( A(t), t = 0, A0 )$
(%i5) atvalue( B(t), t = 0, 0.0 )$
(%i6) atvalue( C(t), t = 0, 0.0 )$

(%i7) desolve( [eqa,eqb,eqc], [A(t),B(t),C(t)] );

Evidently, 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 quick graph of $A(t)$, $B(t)$, and $C(t)$ for a specific set of values with the statements

(%i8) A : rhs( part(%o7,1) )$
(%i9) B : limit(rhs(part(%o7,2)), kA, kB)$
(%i10) C : limit(rhs(part(%o7,3)), kA, kB)$

and plot the graph with the statement

(%i12) plot2d( [A,B,C], [t,0.0,50.0], [xlabel, "t"], [ylabel, "A, B, C"], [legend, false], [color, black], [style, [lines, 4]] );
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.

The resulting graph is shown in Fig. 11.9.

### 11.3.5 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.\(^{18}\)

First, suppressing the display of MAXIMA’s responses, we enter the equations with the statements\(^{19}\)

\[
\begin{align*}
(\%i1) \quad & \text{eq1: } '\text{diff}(x1(t),t,2) + x1(t) - \kappa(x2(t)-x1(t)) = 0 \\
(\%i2) \quad & \text{eq2: } '\text{diff}(x2(t),t,2) + x2(t) + \kappa(x2(t)-x1(t)) = 0
\end{align*}
\]

Then we specify the initial values (again suppressing the display) and invoke \texttt{desolve} with the statements

\[
\begin{align*}
(\%i3) \quad & \text{atvalue( } x1(t), t = 0, x10 ) \\
(\%i4) \quad & \text{atvalue( } x2(t), t = 0, x20 ) \\
(\%i5) \quad & \text{atvalue( } '\text{diff}(x1(t), t), t = 0, 0 ) \\
(\%i6) \quad & \text{atvalue( } '\text{diff}(x2(t), t), t = 0, 0 ) \\
(\%i7) \quad & \text{desolve( [eq1,eq2], [x1(t),x2(t)] )};
\end{align*}
\]

Is \( 2\kappa + 1 \) positive, negative, or zero? p;

---

\(^{18}\)You are urged to explore other initial conditions.

\(^{19}\)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.
Clearly, 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 $\sqrt{2k + 1}$ (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 $\sqrt{2k + 1}$ is present if $x_{20} = -x_{10}$ (the higher-frequency normal mode, in which the objects oscillate with equal amplitude but out of phase).

### 11.3.6 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 would begin by finding the solution to the ODE with the statements

```
(%i1) depends( f, x ) $
(%i2) eq : 'diff( f, x, 2 ) + k^2 * f = 0 $
(%i3) soln : ode2( eq, f, x );
```

Establish dependence of $f$ on $x$. Enter equation. Solve equation for $f$ as function of $x$.

Is $k$ zero or nonzero? n;

```
(%o3) f = %k1 e^{ikx} + %k2 e^{-ikx}
```

Then (naively, as it turns out), we might attempt to impose the boundary conditions with the statement

```
(%i4) bc2(soln, x=0, f=0, x=el, f=0 );
```

but we would quickly find that this statement doesn’t work.\(^{20}\) (MAXIMA returns $f = 0$, having concluded that only the trivial solution to the ODE can be made to satisfy both boundary conditions.) MAXIMA’s routine `bc2` doesn’t recognize that we can find non-trivial solutions, but only if we are prepared to constrain $k$ to a limited set of values. To help MAXIMA recognize that the problem is a bit more subtle, we issue the statements

```
(%i5) eq1 : ev( rhs(soln), x=0 ) = 0; Impose condition at $x = 0$.
(%i5) %k2 + %k1 = 0
(%i6) eq2 : ev( rhs(soln), x=el ) = 0; Impose condition at $x = \ell$.
(%i6) %k1 e^{ik\ell} + %k2 e^{-ik\ell} = 0
```

```
(%i6) %k1 e^{ik\ell} + %k2 e^{-ik\ell} = 0
```

```
(%i7) x1(t) = 1/2 cos(t)(x_{20} + x_{10}) - 1/2 cos(\sqrt{2k + 1} t)(x_{20} - x_{10}),
```

```
(%o7) x2(t) = 1/2 cos(t)(x_{20} + x_{10}) + 1/2 cos(\sqrt{2k + 1} t)(x_{20} - x_{10})
```

Clearly, 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 $\sqrt{2k + 1}$ (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 $\sqrt{2k + 1}$ is present if $x_{20} = -x_{10}$ (the higher-frequency normal mode, in which the objects oscillate with equal amplitude but out of phase).

\(^{20}\) Had we imposed different values on the solution at the two ends of the interval, invocation of `bc2` would have been more successful. Try it, say, with $f(0) = 0$ and $f(\ell) = x_1$. 

The statement at line %i9 also produces a warning, specifically “solve: using arc-trig functions to get a solution. Some solutions will be lost.” MAXIMA fails to recognize that, formally, the solution to \( \sin(k\ell) = 0 \) is \( k\ell = \sin^{-1} 0 \), which admits the solution \( k = n\pi/\ell \), where \( n \) is any positive or negative integer or zero.

Having now found acceptable values of \( k \), we find the corresponding values of \( \%k1 \) and \( \%k2 \) by returning to equations eq1 and eq2. For the allowed values of \( k \),

\[
\begin{align*}
\text{(i10) } & \text{declare( } n, \text{ integer )}$ \\
\text{(i11) } & \text{eq1a : ev( eq1, } k = n*\%pi/\ell \text{ )}$ \\
\text{(i11) } & \text{eq2a : factor( demoivre( } \% \text{ ) )}$
\end{align*}
\]

As expected, the two equations at (i11) and (i12) are the same. Both will be satisfied if we require that \( \%k1 = -\%k2 \), but these equations impose no further constraints. We learn nothing about the remaining constant, which evidently remains arbitrary insofar as the conditions of the problem are concerned. To complete the solution, then, we invoke the statements

\[
\begin{align*}
\text{(i14) } & \text{[\%k1 : -\%k2, } k : n*\%pi/\ell \text{ ]}$ \\
\text{(i15) } & \text{demoivre( ev( rhs(soln )))}$ \\
\text{(i16) } & \text{solna : expand( } \% \text{ )}$ \\
\text{(i17) } & \text{ev(solna, } \%k2 = -A/(2*\%i) \text{ )}$
\end{align*}
\]

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 \( n\pi/\ell \) 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 as simply an alternative choice of \( \%k2 \)). Further, since \( \omega = kc \), we would associate with the solution %o15 the frequency

\[
\omega_n = k_n c = n \frac{\pi c}{\ell} = n \omega_1 \tag{11.70}
\]

where \( \omega_1 \) is the fundamental frequency. We find, in particular, that all allowed frequencies are integer multiples of the fundamental frequency.

### 11.3.7 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
in Section 11.3.6. 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

\[
E_n = \frac{\hbar^2 k_n^2}{2m} = \frac{n^2 \hbar^2 \pi^2}{2m \ell^2}
\]  

(11.71)

Further, we would associate the (unnormalized) wave function

\[
\psi_n(x) = N_n \sin \frac{n\pi x}{\ell}
\]  

(11.72)

with the energy \( E_n \). In the quantum problem, the allowed energies are \textit{not} integer multiples of the lowest energy; rather, the allowed energies increase in proportion to the \textit{squares} of the integers.

### 11.3.8 Finding Solutions by Series Methods

Particularly when analytic methods appear to be yielding little result, we are sometimes tempted to invoke the \textit{method of Frobenius}, representing the solution to the ODE of interest as a power series

\[
f(x) = \sum_{n=0}^{\infty} a_n x^{n+\alpha}
\]

(11.73)

and seeking an \textit{indicial} equation determining \( \gamma \) and \textit{recursion} relationships determining \( a_n \) from a few of the early coefficients. In the absence of a simple MAXIMA 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 MAXIMA 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

(\%i1) \( y : \text{sum( a[i]*x^(i+alpha), i, 0, 8 );} \)

(\%o1) \( a_8 x^{\alpha+8} + a_7 x^{\alpha+7} + a_6 x^{\alpha+6} + a_5 x^{\alpha+5} + a_4 x^{\alpha+4} + a_3 x^{\alpha+3} + a_2 x^{\alpha+2} + a_1 x^{\alpha+1} + a_0 x^{\alpha} \)

(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

(\%i2) \( \text{eq : diff(y,x,2) + y;} \)

(\%o2) \( a_8 x^{\alpha+8} + a_7 x^{\alpha+7} + a_6 (\alpha + 7) x^{\alpha+6} + \ldots + a_1 \alpha (\alpha + 1) x^{\alpha-1} + a_0 \alpha (\alpha - 1) x^{\alpha-2} \)

where we have omitted several of the terms in the middle of this long expression. Not surprisingly, each term has a factor of \( x^\alpha \). Since we ultimately will set this expression equal to zero, we can divide the expression by \( x^\alpha \). To clear the negative powers of \( x \), we also multiply by \( x^2 \). The necessary MAXIMA statements and their output are

(\%i3) \( \text{eq1 : distrib(x^2*eq/x^alpha);} \)

(\%i4) \( \text{eq2 : collectterms( eq1, x^10, x^9, \ldots, x^2, x );} \)

(\%o4) \( a_0 \alpha (\alpha - 1) + a_1 \alpha (\alpha + 1) x + \left[ a_2 (\alpha + 1)(\alpha + 2) + a_0 \right] x^2 + \left[ a_3 (\alpha + 2)(\alpha + 3) + a_1 \right] x^3 + \ldots \)
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 elements of an array and display them with the statement

\[
\text{(\%i5) for } i:0 \text{ thru } 10 \text{ do } \text{coef}[i] : \text{coeff(eq2, x, i) ;}
\]

\[
\text{(\%o5) done}
\]

\[
\text{(\%o6) for } i:0 \text{ thru } 10 \text{ do print("coef[",i,"] =", coef[i] )}
\]

\[
\begin{align*}
\text{coef[0]} &= a_0(\alpha - 1) \\
\text{coef[1]} &= a_1\alpha(\alpha + 1) \\
\text{coef[2]} &= a_2(\alpha + 1)(\alpha + 2) + a_0 \\
\text{coef[3]} &= a_3(\alpha + 2)(\alpha + 3) + a_1 \\
\text{coef[4]} &= a_4(\alpha + 3)(\alpha + 4) + a_2 \\
\text{coef[5]} &= a_5(\alpha + 4)(\alpha + 5) + a_3 \\
\text{coef[6]} &= a_6(\alpha + 5)(\alpha + 6) + a_4 \\
\text{coef[7]} &= a_7(\alpha + 6)(\alpha + 7) + a_5 \\
\text{coef[8]} &= a_8(\alpha + 7)(\alpha + 8) + a_6 \\
\text{coef[9]} &= a_7 \\
\text{coef[10]} &= a_8
\end{align*}
\]

though we must ignore \text{coef[9]} and \text{coef[10]}, which are incomplete because of the truncation of the initial series at \%o1. We then infer that, in the range \( 2 \leq n \leq 8 \) (and presumably beyond 8), the general coefficient satisfies the condition

\[
\text{coef}_n = a_n(\alpha + n - 1)(\alpha + n) + a_{n-2} ; \quad n \geq 2 \quad (11.74)
\]

The statement

\[
\text{(\%i7) coef[n] : a[n]*(alpha+n-1)*(alpha+n) + a[n-2];}
\]

\[
\text{(\%o7) } (n + \alpha - 1)(n + \alpha)a_n + a_{n-2}
\]

will enter this relationship into MAXIMA’s workspace.

Pursuing the solution from this point by the method of Frobenius, we must require (1) \( a_0 \neq 0 \) and (2) the coefficient of each power of \( x \) separately to be zero. For \text{coef[0]} to be zero, we find from the statements

\[
\text{(\%i7) coef[0]/a[0];}
\]

\[
\text{(\%o7) } \alpha(\alpha - 1)
\]

\[
\text{(\%i8) alp : solve( \%=0, alpha );}
\]

\[
\text{(\%o8) } [\alpha = 0, \alpha = 1]
\]

that \( \alpha \) can be either 0 or 1. Then, choosing first \( \alpha = 0 \), we find that \text{coef[1]} = 0 is already satisfied, so we conclude that, with \( \alpha = 0 \), both \( a_0 \) and \( a_1 \) remain undetermined. Finally, to set \( \alpha = 0 \) in Eq. (11.74) and solve for \( a_n, n \geq 2 \), we invoke the MAXIMA statement

\[
\text{(\%i9) part( factor( solve( ev( coef[n], alpha=0 ) = 0, a[n] ) ), 1 ) ;}
\]

\[
\text{(\%o9) } a_n = -\frac{a_{n-2}}{(n-1)n}
\]

to find a recursion relationship that determines all of the coefficients from the first two. Here, \text{factor} replaces \( n^2 - n \) with \( (n - 1)n \) and \text{part} extracts the one element of the one-element list generated by \text{solve}.
We, of course, would like to have explicit values for each of the first several coefficients. We proceed as follows

(%i10) a[n] : -a[n-2]/((n-1)*n)
 (%i11) for i:2 thru 10 do
 a[i] : ev( a[n], alpha=0, n=i )$
 (%i12) y : collectterms( ev( y, alpha=0 ), a[0], a[1] );

Bind solution for a\[n\].
Evaluate coefficients and bind to variables.
Substitute coefficients into original function with \(\alpha = 0\).

\[
\begin{align*}
  a_0 & \left( \frac{x^8}{40320} - \frac{x^6}{720} + \frac{x^4}{24} + \frac{x^2}{2} + 1 \right) + a_1 \left( -\frac{x^7}{5040} + \frac{x^5}{120} - \frac{x^3}{6} + x \right)
\end{align*}
\]

Recognizing that \(2! = 2, 3! = 6, 4! = 24, 5! = 120, 6! = 720, 7! = 5040,\) and \(8! = 40320,\) we then also recognize that the series in the first parentheses is in fact the beginnings of the series for \(\cos(x)\) and the series in the second parentheses is the beginnings of the series for \(\sin(x)\), so the solution to the original equation is, evidently,

\[y(x) = a_0 \cos x + a_1 \sin x\]  \hspace{1cm} (11.75)

We can confirm the series expansions with the MAXIMA statements

(%i13) taylor( cos(x), [x, 0, 8] );
(%o13)/T/ \[
\frac{x^8}{40320} - \frac{x^6}{720} + \frac{x^4}{24} - \frac{x^2}{2} + 1 + \ldots
\]

(%i13) taylor( sin(x), [x, 0, 7] );
(%o13)/T/ \[
-\frac{x^7}{5040} + \frac{x^5}{120} - \frac{x^3}{6} + x + \ldots
\]

The solution when \(\alpha = 1\), which (along with \(\alpha = 0\)) makes \(\text{cof}_0 = 0\) when \(a_0 \neq 0\), can be found in much the same way. In that case, \(\text{cof}_1 = a_1 \alpha (\alpha + 1) = 0\) leads us to conclude that \(a_1 = 0\). When you pursue this alternative to its conclusion, you will find that you have generated only one series, which will turn out to be the series for \(\sin(x)\). We chose \(\alpha = 0\) first because it leads to the full general solution to the original equation, including the two arbitrary constants we know must exist in the general solution to an ordinary, linear, second order ODE. Choosing \(\alpha = 0\) leads only to half of the general solution.

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

\[
\frac{dx}{dt} = f(x, t) \hspace{1cm} x(0) = x_0
\]  \hspace{1cm} (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,\(^{21}\) we can write the approximation

\[
\frac{dx}{dt} = f(x, t) \implies \frac{x(t + \Delta t) - x(t)}{\Delta t} \approx f(x, t) \implies x(t + \Delta t) \approx x(t) + f(x, t) \Delta t \tag{11.77}
\]

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

\[
f_{\text{old}} = f(x_{\text{old}}, t_{\text{old}}) \; ; \; x_{\text{new}} = x_{\text{old}} + f_{\text{old}} \Delta t \; ; \; t_{\text{new}} = t_{\text{old}} + \Delta t \tag{11.78}
\]

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.\(^{22}\)

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 \( f \),

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

\[
A(t + \Delta t) = A(t) + \frac{dA}{dt}(t) \Delta t = A(t) - k_A A(t) \Delta t \tag{11.79}
\]

\[
B(t + \Delta t) = B(t) + \left[k_A A(t) - k_B B(t)\right] \Delta t \tag{11.80}
\]

\[
C(t + \Delta t) = C(t) + k_B B(t) \Delta t \tag{11.81}
\]

\(^{21}\)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 \( \Delta t \ll T \) will probably yield an adequate solution. Each case must be examined on its own terms; no general rules can be formulated.

\(^{22}\)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.
For definiteness, 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 $\Delta t = 0.250$. 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*}

Going yet one more step, Eqs. (11.79), (11.80), and (11.81) for $t = 0.50$ yield that

\begin{align*}
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)
\end{align*}

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.\footnote{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.}

\begin{table}
\centering
\caption{Simple Euler algorithm.}
\begin{tabular}{ll}
\hline
$dt \leftarrow \Delta t$ & Choose time step. \\
$told \leftarrow 0.0$ & Initialize $t_{old}$. \\
xold \leftarrow $x_0$ & Initialize $x_{old}$. \\
\inloop & Evaluate $f_{old}$. \\
\quad fold \leftarrow f(xold, told) & Display results. \\
\quad print, told, xold, fold & Calculate values at new time. \\
\quad exit\_loop when done & Replace old values with new. \\
xnew \leftarrow xold + fold \cdot dt & \\
tnew \leftarrow told + dt & \\
xold \leftarrow xnew & \\
told \leftarrow tnew & \\
\end{tabular}
\end{table}

\footnote{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.}

\footnote{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 $\Delta t$.}
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>1000.000</td>
</tr>
<tr>
<td>0.25</td>
<td>975.000</td>
<td>25.000</td>
<td>0.000</td>
<td>1000.000</td>
</tr>
<tr>
<td>0.50</td>
<td>950.625</td>
<td>48.750</td>
<td>0.625</td>
<td>1000.000</td>
</tr>
<tr>
<td>0.75</td>
<td>926.859</td>
<td>71.297</td>
<td>1.844</td>
<td>1000.000</td>
</tr>
</tbody>
</table>

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
   \[
   x_{\text{pred}} = x_{\text{old}} + f_{\text{old}} \Delta t; \quad t_{\text{new}} = t_{\text{old}} + \Delta t
   \]  
   (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)

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 $f_{\text{old}}$ at the beginning of the interval and (b) the best estimate $f_{\text{pred}}$ 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.\(^{26}\)

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.

\(^{26}\)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. ALGORITHMS FOR SOLVING ODES NUMERICALLY

Table 11.3: Improved Euler algorithm.

```
dt ←− ∆t

told ←− 0.0

xold ←− x₀

loop

fold ←− f(xold, told)

print, told, xold, fold

exit loop when done

xpred ←− xold + fold*dt

tnew ←− told + dt

fpred ←− f(xpred, tnew)

xnew ←− xold + 1/2(fold + fpred)*dt

xold ←− xnew

told ←− tnew

end loop
```

Choose time step.

Initialize $t_{old}$.

Initialize $x_{old}$.

Evaluate $f_{old}$.

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>1000.00</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>

---

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

$$
x(t + \Delta t) = x(t) + \frac{dx(t)}{dt} \Delta t + \frac{1}{2} \left( \frac{d^2x(t)}{dt^2} \right) \Delta t^2 + O(\Delta t^3)
$$

$$
x(t) + f(x(t))\Delta t + \frac{1}{2} \left( \frac{df(x(t), t)}{dt} \right) \Delta t^2 + O(\Delta t^3)
$$

$$
x(t) + f(x(t))\Delta t + \frac{1}{2} \left( \frac{\partial f(x(t), t)}{\partial x} \frac{dx(t)}{dt} + \frac{\partial f(x(t), t)}{\partial t} \right) \Delta t^2 + O(\Delta t^3)
$$

$$
x(t) + f(x(t))\Delta t + \frac{1}{2} \left( f(x(t)) \frac{\partial f(x(t), t)}{\partial x} + \frac{\partial f(x(t), t)}{\partial t} \right) \Delta t^2 + O(\Delta t^3) \quad (11.87)
$$

Motivated by the improved Euler method on the other hand, we are tempted to suppose that we might build a stepping algorithm by\textsuperscript{27}

1. introducing a judicious time $t' = t + a \Delta t$, where $a$, constrained by $0 \leq a \leq 1$, is yet to be specified.

\textsuperscript{27}Note that the steps laid out here reduce to the improved Euler method if we take $a = b = 1$ and $w₁ = w₂ = 1/2$. 
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

\[
f(x', t') = f(x + b \Delta t f(x, t), t + a \Delta t)
\]

\[
= f(x, t) + \frac{\partial f(x, t)}{\partial x} b \Delta t f(x, t) + \frac{\partial f(x, t)}{\partial t} a \Delta t + O(\Delta t^2)
\]

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

\[
x(t + \Delta t) = x(t) + w_1 f(x(t), t) \Delta t
\]

\[
+ w_2 \left( f(x, t) + \frac{\partial f(x, t)}{\partial x} b \Delta t f(x, t) + \frac{\partial f(x, t)}{\partial t} a \Delta t + O(\Delta t^2) \right) \Delta t
\]

\[
= x(t) + (w_1 + w_2) f(x, t) \Delta t
\]

\[
+ \left( w_2 b f(x, t) \frac{\partial f(x, t)}{\partial x} + w_2 a \frac{\partial f(x, t)}{\partial t} \right) \Delta t^2 + O(\Delta t^3)
\]

which will match Eq. (11.87) through terms of order \( O(\Delta t^2) \) if we choose

\[
w_1 + w_2 = 1 \quad ; \quad a w_2 = b w_2 = \frac{1}{2} \quad \text{or} \quad w_1 + w_2 = 1 \quad ; \quad a = b \quad ; \quad w_2 = \frac{1}{2a}
\]

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 27—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

\[
t' = t + \frac{1}{2} \Delta t \quad ; \quad x' = x + \frac{1}{2} f(x, t) \Delta t \quad ; \quad x(t + \Delta t) = x(t) + f(x', t') \Delta t
\]

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

\[\text{footnote text}\]
Table 11.5: A Runge-Kutta Method.

\[ \begin{align*}
\text{dt} & \leftarrow \Delta t \\
t_{\text{old}} & \leftarrow 0.0 \\
x_{\text{old}} & \leftarrow x_0 \\
\text{loop} & \\
\quad \text{fold} & \leftarrow f(x_{\text{old}}, t_{\text{old}}) \\
\quad \text{print, } t_{\text{old}}, x_{\text{old}}, \text{ fold} \\
\quad \text{exit loop when done} \\
\quad k_1 & \leftarrow \text{fold} \times \text{dt} \\
\quad k_2 & \leftarrow f(x_{\text{old}} + k_1/2, t_{\text{old}} + \text{dt}/2) \times \text{dt} \\
\quad t_{\text{new}} & \leftarrow t_{\text{old}} + \text{dt} \\
\quad x_{\text{new}} & \leftarrow x_{\text{old}} + (k_1 + k_2)/2 \\
x_{\text{old}} & \leftarrow x_{\text{new}} \\
t_{\text{old}} & \leftarrow t_{\text{new}} \\
\text{end loop}
\end{align*} \]

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 \( k_1, k_2, \) and \( x_{\text{new}} \) in the above algorithm:

\[ \begin{align*}
\text{k}1 & \leftarrow f(x_{\text{old}}, t_{\text{old}}) \times \text{dt} \quad \text{Find change using rate at beginning.} \\
\text{k}2 & \leftarrow f(x_{\text{old}} + \text{k}1/2, t_{\text{old}} + \text{dt}/2) \times \text{dt} \quad \text{Estimate change using rate at midpoint.} \\
\text{k}3 & \leftarrow f(x_{\text{old}} + \text{k}2/2, t_{\text{old}} + \text{dt}/2) \times \text{dt} \quad \text{Estimate change using refined rate at midpoint.} \\
\text{k}4 & \leftarrow f(x_{\text{old}} + \text{k}3, t_{\text{old}} + \text{dt}) \times \text{dt} \quad \text{Estimate change using rate at end.} \\
x_{\text{new}} & \leftarrow x_{\text{old}} + (\text{k}1 + 2\times\text{k}2 + 2\times\text{k}3 + \text{k}4)/6 \quad \text{Calculate final value.}
\end{align*} \]

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 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. \textit{The importance of being aware that numerical methods are always approximate cannot be overstressed.}

Two distinctly different sorts of errors can occur. \textit{Truncation errors} arise because the solution has been based on a finite-difference approximation to the derivatives appearing in the equation; \textit{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

\[ \left| x_{\text{exact}}(t + \Delta t) - x_{\text{Euler}}(t + \Delta t) \right| \leq \text{Max} \left( \left| \frac{1}{2} \frac{d^2 x}{dt^2}(\xi) \right| \right) \Delta t^2 \]  

(11.94)

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

\[ x_{\text{exact}}(t + \Delta t) = x_{\text{Euler}}(t + \Delta t) + O(\Delta t^2) \]  

(11.95)

\[ = x_{\text{ImpEul}}(t + \Delta t) + O(\Delta t^3) \]  

(11.96)

\[ = x_{\text{RK2}}(t + \Delta t) + O(\Delta t^4) \]  

(11.97)

\[ = x_{\text{RK4}}(t + \Delta t) + O(\Delta t^5) \]  

(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 \( \Delta t \), \( \Delta t^2 \), \( \Delta t^2 \), 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

---

\(^{29}\) which provides yet another motivation for casting problems in dimensionless form.

\(^{30}\) This assertion is refined in the next paragraphs.

\(^{31}\) This approach is not entirely secure if the procedure converges slowly as the time step is reduced. Remember that \( \sum_{n=1}^{\infty} \frac{1}{n} \) diverges even though the effect of the millionth term added to the partial sum starts in the sixth decimal place.

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 \textit{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 \textit{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 \( NO(\Delta t^p) \) 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 \textit{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) \textit{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.\(^{33}\) 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 \textit{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.\(^{34}\)

---

\(^{33}\)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.

\(^{34}\)In contrast, Euler’s method, for example, would require using throughout a time step small enough to generate
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 $x_i = x(t_i)$ 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

$$x_{i+1} = x_i + \frac{\Delta t}{24} (55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3}) \quad (11.99)$$

(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 previous 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.

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

$$x_{i+1} = x_i + \frac{\Delta t}{24} (9f_{i+1} + 19f_i - 5f_{i-1} + f_{i-2}) \quad (11.100)$$

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.

11.8 Solving ODEs Numerically with MATLAB

Note: All MATLAB program (.m) files referred to in this chapter are available in the directory $\text{HEAD}/\text{matlab}$, where (as defined in the Local Guide) $\text{HEAD}$ 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 MATLAB’s default search path. If so, the files can be found by MATLAB 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 MATLAB’s elementary commands or, more simply, by invoking one of the built-in routines, which include ode23, ode45, and ode113.

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.

11.8. SOLVING ODES NUMERICALLY WITH MATLAB

11.8.1 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 \(0 \leq t \leq 50.0\) with a time step of \(\Delta t = 0.25\) and \(k_A = k_B = 0.1\). We could exploit the capabilities of MATLAB 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

```matlab
>> t(1)=0.0; A(1)=1000.0;
>> B(1)=0.0; C(1)=0.0;
>> kA = 0.1; kB = 0.1;
>> dt = 0.25;
```

Set initial values.

Set parameters.

Set time step.

The solution is then calculated, time instant by time instant, with the loop

```matlab
>> 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 \(t\).\(^{36}\) The simple statement

```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
```

displays the resulting solution\(^{37}\) 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
```

\(^{36}\)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 \(t(i) = (i-1) \times dt\), though doing so would increase the execution time. (Multiplication takes more time than addition.)

\(^{37}\)We have removed several extraneous lines from the output produced by this statement.
CHAPTER 11. SOLVING ORDINARY DIFFERENTIAL EQUATIONS

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.

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>6.0000</td>
<td>2.0000</td>
<td>512.000</td>
<td>384.000</td>
<td>104.000</td>
</tr>
<tr>
<td>1.0000</td>
<td>531.441</td>
<td>354.294</td>
<td>114.265</td>
<td></td>
</tr>
<tr>
<td>0.5000</td>
<td>540.360</td>
<td>341.280</td>
<td>118.360</td>
<td></td>
</tr>
<tr>
<td>0.2500</td>
<td>544.642</td>
<td>335.164</td>
<td>120.194</td>
<td></td>
</tr>
<tr>
<td>0.1250</td>
<td>546.740</td>
<td>332.197</td>
<td>121.063</td>
<td></td>
</tr>
<tr>
<td>0.0625</td>
<td>547.779</td>
<td>330.735</td>
<td>121.486</td>
<td></td>
</tr>
</tbody>
</table>

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 2 is apparent. 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 $\pm 0.1$ 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 $\pm 0.1$ 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
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.

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>551.368</td>
<td>322.752</td>
<td>125.880</td>
</tr>
<tr>
<td>1.0000</td>
<td>549.404</td>
<td>327.821</td>
<td>122.776</td>
<td></td>
</tr>
<tr>
<td>0.5000</td>
<td>548.954</td>
<td>328.940</td>
<td>122.106</td>
<td></td>
</tr>
<tr>
<td>0.2500</td>
<td>548.846</td>
<td>329.202</td>
<td>121.951</td>
<td></td>
</tr>
<tr>
<td>0.1250</td>
<td>548.820</td>
<td>329.266</td>
<td>121.914</td>
<td></td>
</tr>
<tr>
<td>0.0625</td>
<td>548.814</td>
<td>329.282</td>
<td>121.904</td>
<td></td>
</tr>
</tbody>
</table>

converging more quickly on stable values to more decimal places than we observed in Table 11.6. 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.
Table 11.8: A generic function to define ODEs for MATLAB.

```matlab
function Derivs = FunctionName( IndVar, DepVars, P1, P2, ... )
    \% Single descriptive help line.
    \% Explanatory comments.
    ...
    Intermediate statements.
    ...
    Derivs = [ DerivOfFirstVar; DerivOfSecondVar; ... ]
end
```

### 11.8.2 Defining ODEs for MATLAB

Before using any of the routines MATLAB supplies for solving a system of first-order differential equations numerically, we must create an M-file\(^{38}\) whose execution by MATLAB 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 the most recent versions of MATLAB,\(^{39}\) this file will—in broad outline—have the form shown in Table 11.8. In this structure,

- 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` (mandatory) is the name of the variable in which the function will return the derivatives of the dependent variables; the returned result must be a column vector.
- `IndVar` (mandatory) is a scalar that, on input, supplies the value of the independent variable.
- `DepVars` (mandatory) is a vector that, on input, supplies the values of the dependent variables.
- `P1, P2, ...` (optional) on input supply values of any parameters that may be needed to calculate the derivatives. If the differential equations of interest have no internal parameters, these arguments will simply be omitted in the definition of the function.

Explicit M-files for several examples will be presented in the subsequent subsections.

### 11.8.3 The Procedures ode23, ode45, and ode113

MATLAB contains several built-in procedures for solving ODEs. All are invoked with the same syntax, here illustrated with the procedure `ode23`, which is adaptive and uses a combination of second and third-order Runge-Kutta methods to step the solution forward by a specified amount, holding the solution to an absolute (truncation) error per step no larger than \(10^{-6}\) in absolute value or \(10^{-3}\) in relative value (whichever is larger).\(^{40}\) A statement invoking `ode23` has the general form

\[
[T, Soln] = \text{ode23}( \text{@FunctionName}, \text{TimeInt}, \text{ICs}, \text{Options}, \text{Param1}, \text{Param2}, \ldots )
\]

\(^{38}\)See Section 3.9.

\(^{39}\)The syntax here described was introduced no more recently than MATLAB 2012a. Note that differences exist only if optional parameters were needed. In the absence of such parameters, the newer syntax and the older syntax were the same.

\(^{40}\)See Section 11.8.9 for the ways these tolerances can be changed.
11.8. SOLVING ODES NUMERICALLY WITH MATLAB

In this call to \texttt{ode23},

- \( T \) (mandatory) is the name of the variable—ultimately a vector—that will store the times at which solutions are generated,
- \( \text{Soln} \) (mandatory) is the name of the variable—ultimately a matrix with one column for each dependent variable—that will store the solutions generated,
- \( \text{FunctionName} \) (mandatory) is the name of the function that returns the derivatives,
- \( \text{TimeInt} \) (mandatory) is a two-component vector giving the starting and ending times for the solution (which will then be generated at unequally spaced times) or a multi-component vector giving the times—equally or unequally spaced—at which a solution is to be returned,
- \( \text{ICs} \) (mandatory) is a \textit{column} vector whose elements give the initial values of the dependent variables,
- \( \text{Options} \) (optional) specifies options and, for our purposes will usually be null (conveyed by \texttt{[]} if it is needed at all), and
- \( P1, P2, \ldots \) (optional) specify parameters that will be passed on to \text{FunctionName} at the time of execution.

The second built-in procedure for solving ODEs is \texttt{ode45}, an adaptive procedure whose arguments are the same as those of \texttt{ode23} but that uses a combination of fourth and fifth-order Runge-Kutta techniques. It is therefore a higher order method than that implemented in \texttt{ode23} and, for a given time step will produce more accurate solutions than \texttt{ode23}. Equivalently, \texttt{ode45} will achieve a given accuracy with larger time steps than will be required by \texttt{ode23}.

The solver \texttt{ode113} is a multistep solver that uses the Adams-Bashforth-Moulton method. Its use is similar to that of \texttt{ode23} and \texttt{ode45}; those interested in more detail are referred to the MATLAB manuals.

The procedures \texttt{ode23}, \texttt{ode45}, and \texttt{ode113} are satisfactory for all of the exercises we will encounter in this book. When it happens that the solution changes very rapidly over a short time interval and we seek a solution over a much longer time interval, however, more sophisticated methods may be necessary. Under those circumstances, the equation(s) is (are) said to be \textit{stiff}, and the simpler algorithms embodied in \texttt{ode23}, \texttt{ode45}, and \texttt{ode113} require embellishment. For these circumstances, MATLAB provides the routines \texttt{ode15s}, \texttt{ode23s}, \texttt{ode23t}, and \texttt{ode23tb}, which are described in the MATLAB manuals.

11.8.4 Chain Radioactive Decay

As a first example of the use of a MATLAB routine, consider the equations describing chain radioactive decay, Eq. (11.19). Before we can construct the proper M-file to communicate the differential equations to MATLAB’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 MATLAB 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 \( A \mapsto n(1), B \mapsto n(2), \) and \( C \mapsto n(3) \). Rewriting the differential equations Eq. (11.19) in terms of the elements of \( n \), we have the system of equations

\[
\frac{dn(1)}{dt} = -k_A n(1) \quad ; \quad \frac{dn(2)}{dt} = k_A n(1) - k_B n(2) \quad ; \quad \frac{dn(3)}{dt} = k_B n(2) \quad (11.101)
\]
function derivs = decay( t, n, kA, kB )
% 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)
derivs = [-kA*n(1); kA*n(1)-kB*n(2); kB*n(2)]
end

Now we are ready to construct the function M-file describing this system of differential equations.
Respecting the general structure described in Section 11.8.2, we would compose the definition in Table 11.9. Here, the first line tells MATLAB that the function returns the derivatives in the variable derivs (a vector), that the function is named decay, that the independent and dependent variables will be supplied in t—a scalar—and n—a vector, and that kA and kB will communicate the values of the parameters. The last line (which is the only significant line in the file) assigns to the value of the function a column vector whose elements are the derivatives of the dependent variables as determined by the right hand sides of Eqs. (11.101).

If we store the function of the previous paragraph in the file decay.m in the current directory, we would produce a solution for the differential equations it defines with the statements

```matlab
>> kA = 0.1; kB = 0.1;
>> ic = [1000.0; 0.0; 0.0 ];
>> [ t, n ] = ode23( @decay, [0.0 50.0], ...
    ic, [], kA, kB );
```

Then we would plot the three dependent variables on the same axes with the statements

```matlab
>> plot( t, n(:,1), 'Color', 'black', 'LineWidth', 2.0 )
>> hold on
>> plot( t, n(:,2), 'Color', 'black', 'LineWidth', 2.0, 'LineStyle', ':' )
>> plot( t, n(:,3), 'Color', 'black', 'LineWidth', 2.0, 'LineStyle', '--' )
>> title( 'Three-Species Chain Decay', 'FontSize', 16 )
>> xlabel( 'Time', 'FontSize', 14 )
>> ylabel( 'Number of Atoms Present', 'FontSize', 14 )
>> set( gca, 'FontSize', 14 )
>> hold off
```

The resulting graph—Fig. 11.12—shows the time variation of the number of atoms of each species.

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 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 the second/third order Runge-Kutta method of ode23, with the (default) tolerance of $10^{-6}$ absolute, $10^{-3}$ relative (see Section 11.8.9), and was accomplished with 37 points, as determined by executing the MATLAB command

```matlab
>> ode23( @decay, [0.0 50.0], ...
    ic, [], kA, kB );
```
11.8. SOLVING ODES NUMERICALLY WITH MATLAB

Figure 11.12: Solution of chain radioactive decay via ode23 with a tolerance of $10^{-6}$ absolute, $10^{-3}$ relative (see Section 11.8.9). The solid, dotted and dashed lines show $A(t)$, $B(t)$, and $C(t)$, respectively.

Let us also generate a solution to this same problem using ode45. The function M-file decay.m will serve in this context as well. Except for replacing ode23 with ode45, we would invoke statements identical to those invoked above, namely

```matlab
>> kA = 0.1; kB = 0.1;
>> ic = [1000.0; 0.0; 0.0 ];
>> [ t, n ] = ode45( @decay, [0.0 50.0], ...
ic, [], kA, kB);
```

A graph of the resulting solution would be produced by statements identical to those that produced Fig. 11.12, and the resulting graph is essentially identical to Fig. 11.12.
CHAPTER 11. SOLVING ORDINARY DIFFERENTIAL EQUATIONS

Table 11.10: The MATLAB function damposc.m.

function derivs = damposc( t, y, bb )
% 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.
    derivs = [ y(2); -y(1) - bb*y(2) ];
end

11.8.5 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 \( f(t) \) 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

\[
\frac{d}{dt}y(1) = y(2) ; \quad \frac{d}{dt}y(2) = -y(1) - \beta y(2)
\]  

(11.102)

Then, using the symbol \( bb \) for \( \beta \), we define the differential equations by creating the M-file in
Table 11.10. 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
>> bb=0.0;
>> ic=[ 1.0; 0.0 ];
>> [t1, x1] = ode23( @damposc, [0.0 30.0], ...
    ic, [], bb );
>> bb=0.25;
>> [t2, x2] = ode23( @damposc, [0.0 30.0], ...
    ic, [], bb );
```

Finally, we produce graphs showing the position and velocity as functions of time and also the
trajectory in the phase plane \( (v \text{ versus } x) \) for each value of the damping constant with the statements

```matlab
>> subplot(2,2,1)
>> plot( t1, x1(:,1), 'LineWidth', 2.0, 'Color', 'black' )
>> title( 'Undamped Harmonic Oscillator' )
>> xlabel('Time')
>> hold on
>> plot( t1, x1(:,2), 'LineWidth', 2.0, 'Color', 'black', 'LineStyle', '--' )
>> hold off

>> subplot(2,2,2)
>> plot( t2, x2(:,1), 'LineWidth', 2.0, 'Color', 'black' )
>> title( 'Damped Harmonic Oscillator' )
>> xlabel('Time')
>> hold on
>> plot( t2, x2(:,2), 'LineWidth', 2.0, 'Color', 'black', 'LineStyle', '--' )
```
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.

```matlab
>> hold off
>> subplot(2,2,3)
>> plot( x1(:,1), x1(:,2), 'LineWidth', 2.0, 'Color', 'black' )
>> xlabel( 'Position' ); ylabel( 'Velocity' )
>> subplot(2,2,4)
>> plot( x2(:,1), x2(:,2), 'LineWidth', 2.0, 'Color', 'black' )
>> xlabel( 'Position' ); ylabel( 'Velocity' )
```

The resulting graphs are shown in Fig. 11.13.

Alternatively, we could, of course, approach this problem using `ode45`. The statements differ from those earlier in this section simply by the replacement of `ode23` with `ode45` and graphs essentially identical to those in Fig. 11.13 would be produced.

### 11.8.6 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
function derivs = planet( t, x, b )
    % 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.
    temp=( x(1)^2 + x(3)^2 )^b;
    derivs = [ x(2); -x(1)/temp; x(4); -x(3)/temp ];
end

The system
\[
\frac{dx}{dt} = \nu_x ; \quad \frac{d\nu_x}{dt} = -\frac{x}{(x^2 + y^2)^b} ; \quad \frac{dy}{dt} = \nu_y ; \quad \frac{d\nu_y}{dt} = -\frac{y}{(x^2 + y^2)^b} \quad (11.103)
\]

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 \( x, \nu_x, y, \) and \( \nu_y \) 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 \( \text{planet.m} \) defining the above equations is shown in Table 11.11. Finally, to calibrate our sense of appropriate initial conditions, we remember from Eq. (11.50) 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 \( 1/\sqrt{4.0} = 0.5 \) units.

Now we have all the elements in place to invoke MATLAB and pursue solutions to the planetary problem. Electing to use \( \text{ode45} \) and accepting its default tolerances (see Section 11.8.9), 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

```matlab
>> b=1.5;
>> ic1=[ 4.0; 0.0; 0.0; 0.5 ];
>> [ t1, x1 ] = ode45( @planet, [0.0 60.0], ic1, [], b );
>> ic2=[ 4.0; 0.0; 0.0; 0.3 ];
>> [ t2, x2 ] = ode45( @planet, [0.0 60.0], ic2, [], b );
```

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.\(^{41}\)

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

```matlab
>> E1 = 0.5*(x1(:,2).^2+x1(:,4).^2) - 1.0./sqrt(x1(:,1).^2+x1(:,3).^2);
```
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 11.12: MATLAB statements for open orbits.

```matlab
>> b=1.5;
>> axis square;
>> set( gca, 'XLim', [-4.0,4.0], 'YLim', [-4.0,4.0], 'FontSize', 14 );
>> title( 'Open Orbits', 'FontSize', 16 )
>> xlabel( 'X-Position', 'FontSize', 14 )
>> ylabel( 'Y-Position', 'FontSize', 14 )
>> hold on
>> for i= -4.0:4.0
    ic=[-4.0; 1.0; i; 0.0];
    [t, x] = ode45( @planet, [0.0 30.0], ic, [], b );
    plot( x(:,1), x(:,3), 'LineWidth', 2.0, 'Color', 'black' );
end
>> grid on;
>> 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);
```

will calculate these quantities. Then, statements like

```matlab
>> E1, L1
```

will display the resulting values. We find—see your own output—that \( E_1 \) = \(-0.1250 \) at the beginning and climbs to \(-0.1266 \) by the end of the interval while \( L_1 \) starts at 2.000 and falls to 1.9873; \( E_2 \) varies between \(-0.2050 \) and \(-0.2069 \) while \( L_2 \) varies between 1.2032 and 1.1992 throughout the interval of solution. 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 MATLAB’s `for` loop to produce a variety of open orbits for a planet started in different initial positions by executing the statements listed in Table 11.12. Figure 11.15 shows the resulting output.

11.8.7 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 **shooting method**) involves accepting the boundary value at one end of the interval, **guessing** a derivative at that same end, assuming values for any parameters in the equation, solving the resulting 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

\[
\frac{df}{dx} = g \quad ; \quad \frac{dg}{dx} = -k^2 f \quad ; \quad f(0) = 0 \quad ; \quad f(\ell) = 0
\]  

(11.104)

Thus, setting up the correspondences \( f \mapsto f(1) \) and \( g \mapsto f(2) \), we might define this system for MATLAB with the M-file shown in Table 11.13 and stored in the default directory with the name
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). Choosing ode23, we would invoke the statements

```matlab
>> k = 1.0;    % Set parameter.
>> ic = [0.0; 1.0]; % Set initial conditions.
>> [ x, f] = ode23( @stdwaves, [0.0 1.0],...  % Invoke solver.
                ic, [], k );
>> plot( x, f(:,1), 'LineWidth', 2.0, 'Color', 'black' );
>> set( gca, 'YLim', [-1.0, 1.0] );    % Plot solution.
>> grid on;
```

The result is the higher solid line in Fig. 11.16. Repeating this calculation with the more accurate

---

**Table 11.13: The MATLAB function stdwaves.m.**

```
function derivs = stdwaves( x, f, k )
% STD WAVES: returns derivatives for standing waves in string
% The function STD WAVES defines the basic equations for describing
% standing waves in a string.

derivs = [ f(2); -k^2*f(1) ];
end
```

---

Figure 11.15: Open orbits of a satellite moving under the influence of a sun at the origin.
Figure 11.16: Shooting method to solve for standing waves in a string. The upper solid, dotted, dashed, and lower solid curves correspond to $k = 1.0, 2.0, 3.0, \text{and} 4.0$, respectively.

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

\begin{verbatim}
>> hold on;
>> k = 2.0;
>> [ x, f ] = ode23( @stdwaves, [0.0 1.0], ic, [], k );
>> plot( x, f(:,1), 'LineWidth', 2.0, 'Color', 'black', 'LineStyle', ':' );
>> k = 3.0;
>> [ x, f ] = ode23( @stdwaves, [0.0 1.0], ic, [], k );
>> plot( x, f(:,1), 'LineWidth', 2.0, 'Color', 'black', 'LineStyle', '--' );
>> k = 4.0;
>> [ x, f ] = ode23( @stdwaves, [0.0 1.0], ic, [], k );
>> plot( x, f(:,1), 'LineWidth', 2.0, 'Color', 'black', 'LineStyle', '-' );
>> hold off
\end{verbatim}

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

\[
>> k = 0:0.2:100.0/5.0;
>> for i = 1:101
    [t, f] = ode23( @stdwaves, [0.0 1.0], ic, [], k(i) );
    q = whos('t'); soln(i) = f( q.size(1), 1 );
end
>> plot( k, soln, 'LineWidth', 2.0, 'Color', 'black' )
>> grid on
\]

Here, we (1) establish values for \( k \), (2) execute a loop that (a) invokes \texttt{ode23} to solve the equations, (b) uses \texttt{whos} to determine the number of entries in the arrays containing the solution,\(^{42}\) and (c) stores the final value (which will be in row \( q.size(1) \), and (3) 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 \( 0 \leq k \leq 20 \).

### 11.8.8 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 earlier, we can generate a more tractable 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\(^{43}\) \( 0 \leq y < \infty \). 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 \texttt{qmshm.m} defining these equations for MATLAB’s solvers is shown in Table 11.14. With this file stored in the default directory with the name \texttt{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,

\(^{42}\)Used in this way, the function \texttt{whos} returns a structured variable containing several elements. The element \( q.size \) is a two-component vector whose first component gives the number of elements in the variable \( t \). We must adopt this route because \texttt{ode23} returns a solution with irregular time steps and the vectors containing the solutions for the several values of \( k \) may not all have the same number of elements.

\(^{43}\)We here use the symbol \( y \) for what we earlier called \( x \).
Figure 11.17: Graph of $f(1)$ versus $k$ for standing waves in a string.

Table 11.14: The MATLAB function \texttt{qmshm.m}.

```matlab
function derivs = qmshm( y, psi, ee )
% QMSHM: returns the derivatives for the quantum harmonic oscillator
% QMSHM defines the dimensionless Schroedinger equation for the quantum
% simple harmonic oscillator, which is
% \[
% \frac{d^2 \psi}{dy^2} + (2 \epsilon - y^2) \psi = 0
% \]
% Entries are psi and d psi/dy, and initial conditions will thus be
% [psi0, dpsi0]. The parameter ee represents a dimensionless energy.

derivs = [ psi(2); -(2*ee-y^2)*psi(1) ];
end
```

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
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

```matlab
>> e = [ 0.3 0.4 0.5 0.6 0.7 ];
>> q = size(e);
>> [y, psi] = ode23( @qmshm, [0.0 4.0], ic, [], e(1) );
>> 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)
    [y, psi] = ode23( @qmshm, [0.0 4.0], ic, [], e(i) );
    plot( y, psi(:,1), 'LineWidth', 2.0, 'Color', 'black' );
end
>> plot( [0.0,4.0], [0.0, 0.0], 'Color', 'black' );
>> text( 2.0, 1.7, '0.3' );
>> text( 2.8, 1.2, '0.4' );
>> text( 3.5, 0.1, '0.5' );
>> text( 2.8, -1.0, '0.6' );
>> text( 2.2, -1.6, '0.7' );
>> hold off;
```

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' );
>> text( 1.7, -1.7, '1.5' );
>> text( 2.7, -1.8, '2.0' );
>> text( 2.7, -0.3, '2.5' );
>> text( 2.7, -1.8, '2.0' );
>> text( 2.7, -1.8, '2.0' );
>> text( 3.0, 1.2, '3.0' );
>> text( 2.5, 1.7, '3.5' );
```

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
Figure 11.18: The solution to the Schrödinger equation for an even state of the quantum harmonic oscillator when $\epsilon$ has the indicated values. This state is the ground state (or state of lowest energy).

Figure 11.19: The solution to the Schrödinger equation for an even state of the quantum harmonic oscillator when $\epsilon$ has the indicated values. As we shall confirm in the next figure, this state is actually the second excited state of the harmonic oscillator.
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

```matlab
>> ic = [ 0.0; 1.0 ];
>> e = [ 1.3 1.4 1.5 1.6 1.7 ];
```

Then we execute the statements

```matlab
>> q = size(e);
>> [y, psi] = ode23( @qmshm, [0.0 4.0], ic, [], e(1) );
>> plot( y, psi(:,1), 'LineWidth', 2.0, 'Color', 'black' );
>> axis( [0.0 4.0 -2.0 2.0 ] );
>> hold on;
>> for i = 2:q.size(2)
>>   [y, psi] = ode23( @qmshm, [0.0 4.0], ic, [], e(i) );
>>   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

```matlab
>> text( 2.7, 1.7, '1.3' );
>> text( 3.5, 1.5, '1.4' );
>> text( 3.5, 0.1, '1.5' );
>> text( 3.4, -1.0, '1.6' );
>> text( 2.8, -1.5, '1.7' );
>> 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

$$E_n = \left(n + \frac{1}{2}\right) \hbar \omega$$

(11.106)

though we have direct evidence for this (correct) conclusion only for the lowest three states.

### 11.8.9 Changing Tolerances

The detailed behavior of MATLAB’s ODE solvers as a group is controlled by a number of properties, 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. The command `odeset` without arguments will display a list of all of these properties together with an indication of the default value and the allowed values. Among the more important of these properties are
Figure 11.20: The solution to the Schrödinger equation for an odd state of the quantum harmonic oscillator when $\epsilon$ has the indicated values.

- **AbsTol** (default $10^{-6}$), 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.

- **RelTol** (default $10^{-3}$), which specifies the maximum relative error in the dependent variables. This property can be only a scalar, and it applies to all dependent variables.

- **InitialStep**, which specifies a starting value for the step size, which is then adjusted as the adaptive procedure works out the solution. If no value is specified, MATLAB makes its own choice.

- **MaxStep**, which specifies the largest step size the solver is to use. If no value is specified, MATLAB makes its own choice.

The two tolerances **AbsTol** and **RelTol** are both taken into account as MATLAB assesses convergence. Specifically, MATLAB seeks to keep the absolute value of the error in the dependent variable no bigger than the quantity

$$\max(\text{RelTol}\cdot\text{abs}(y), \text{AbsTol})$$

Thus, when the solution is close to zero, the absolute tolerance provides the controlling criterion while for solutions 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.

To this point, we have accepted the default tolerances for each of MATLAB’s ODE solvers. The values can, however, be changed by exploiting the fourth argument to the solver—the argument that we have heretofore routinely set to a null string. If, in the solution of the problem in radioactive
11.8. SOLVING ODES NUMERICALLY WITH MATLAB

decay, we had wanted to change the absolute tolerance to $10^{-9}$ and the relative tolerance to $10^{-6}$, we would have used the statements

```matlab
>> opt = odeset('AbsTol', 1.0e-9, ... Define opt with new values.
                  'RelTol', 1.0e-6 );
```

Set parameters.

```matlab
>> kA = 0.1; kB = 0.1; ic = [1000.0; 0.0; 0.0 ];
```

Set initial conditions.

```matlab
>> [ t, n ] = ode23(@decay, [0.0 50.0], ...
                    ic, opt, kA, kB );
```

Solve the system with the options in opt (and all others at their default values).

In contrast to our previous solution, which generated vectors \( t \) and \( n \) with 37 elements, this time these vectors contain 289 elements; the solution has been generated to a higher accuracy.

Full detail about all of the properties that affect the behavior of MATLAB's ODE solvers can be found in the MATLAB manuals.

11.8.10 M-files in the Public Library

The following M-files are found in the directory \$HEAD/matlab (and may also at some sites have been placed in a directory in MATLAB'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

```matlab
>> help $HEAD/matlab/decay
```

to MATLAB or with whatever command in your operating system lists the file on the screen.\footnote{In UNIX, one possible command would be \texttt{more \$HEAD/matlab/decay.m}; see the \textit{Local Guide}.} The first of these commands instructs MATLAB 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 will display the entire file on the screen. Many of the functions defined by these files involve parameters that must be supplied among the arguments in a call to an ODE solver; examine the help message and/or the file itself to determine the number and function of any parameters.

- **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.
- **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**: describes the behavior of standing waves in a string.

• **twodshm.m**: describes the two-dimensional, undamped, unforced harmonic oscillator using dimensionless variables.

• **twooscil.m**: describes a system of two equal masses between two walls, all connected by three springs. Dimensionless units are used.

• **vandpol.m**: describes the Van der Pol oscillator.

### 11.11 Solving ODEs Numerically with MAXIMA

Numerical approaches to solving ODEs most commonly expect that the equation(s) to be solved will be presented as a system of first-order equations of the form

\[
\frac{dx_i}{dt} = f(x_1, x_2, \ldots, x_n, t), \quad 1 \leq i \leq n
\]  

(11.107)

The one command available in standard MAXIMA for solving ordinary differential equations numerically is `rk`, which uses the fourth-order Runge-Kutta method and has the general syntax

\[
rk( \text{Deriv(s)}, \text{DepVar(s)}, \text{InitVal(s)}, [\text{IndepVar}, \text{Start}, \text{Stop}, \text{Incr}] );
\]

where `Deriv(s)` is the derivative of the single dependent variable or a list of such derivatives, `DepVar(s)` is the dependent variable or a list of those variables, `InitVal(s)` is the initial condition or a list of those conditions, and—in a list with four components—`IndepVar` is the (single) independent variable, `Start` and `Stop` are the starting and stopping values of the independent variable, and `Incr` specifies the increment by which the independent variable is to be stepped from `Start` to `Stop`. In this case, both the differential equations and the initial conditions must be specified; MAXIMA will not introduce undetermined integration constants. Further, any parameters in the equations must, of course, be given explicit numerical values before `rk` is executed. When invoked, `rk` returns a solution expressed as a list of lists, with each of the interior lists being a list whose first element is the value of the independent variable and whose remaining elements are the solution(s) to the differential equation(s) in the order in which the dependent variable(s) is (are) presented in the second argument of `rk`. MAXIMA’s `part` command can be used to extract the solution at a single value of the independent variable.

Thus, for example, to address the problem of chain radioactive decay that we have already solved symbolically in Section 11.3.4, we would first define the equations and initial conditions as lists with the statements

\[
(\%i1) \text{eqs} : [ -kA*A, kA*A -kB*B, kB*B ]$
\]

\[
(\%i2) \text{ics} : [ 1000, 0, 0 ]$
\]

Then we would give explicit numerical values to the parameters with the statements

\[\text{45 Higher-order equations must be recast as a system of first-order equations.}\]
Next, we would invoke the solver with the statement

(%i5) soln : rk( eqs, [A,B,C], ics, [t, 0, 50, 1 ] )$

Here, we have suppressed the output because it is a bit of a mess, consisting of a list of 51 four-component sublists with most components being double-precision values. Each of the four-component sublists has the structure \([t, a(t), b(t), c(t)]\). To illustrate more fully, we extract the first, second, and last four-component sublists with the statements

(%i6) part( soln, 1 );
(%o6) \[0.0, 1000.0, 0.0, 0.0\]
e.g., at \(t = 0.0\), \(A(t) = 1000.0, B(t) = 0.0, C(t) = 0.0\)

(%i7) part( soln, 2 );
(%o7) \[1.0, 904.8375, 90.483333333333, 4.67916666666667\]
e.g., at \(t = 1.0\), \(A(t) = 904.83\ldots, B(t) = 90.48\ldots, C(t) = 4.67\ldots\)

(%i8) part( soln, 51 );
(%o8) \[50.0, 6.737977516754986, 33.68973244593903, 959.5722900373056\]
e.g., at \(t = 50.0\), \(A(t) = 6.73\ldots, B(t) = 33.68\ldots, C(t) = 959.57\ldots\)

A graphical display of this solution is most useful. To exploit \texttt{plot2d}, however, we need separate lists of the pairs \(AA : [t, A(t)], BB : [t, B(t)],\) and \(CC : [t, C(t)]\), which we extract from \texttt{soln} with the statements

(%i9) AA : makelist( [ p[1],p[2] ], p, soln );
(%o9) \[[0.0, 1000.0],[1.0, 904.83\ldots],[50.0, 6.73\ldots]\]

(%i10) BB : makelist( [ p[1],p[3] ], p, soln );
(%o10) \[[0.0, 0.0],[1.0, 90.48\ldots],[50.0, 33.68\ldots]\]

(%i11) CC : makelist( [ p[1],p[4] ], p, soln );
(%o11) \[[0.0, 0.0],[1.0, 4.67\ldots],[50.0, 959.57\ldots]\]

Here, \(p\) is a dummy variable (and can be anything). \texttt{makelist} has three arguments, and the first statement, for example, assigns to the variable \(AA\) a list of two-component (sub)lists, each of which contains the first and second components of the corresponding four-component (sub)list in \texttt{soln}. With these variables in hand, we then create the desired graph with the statement

\texttt{plot2d( [discrete,AA], [discrete,BB], [discrete,CC] ), [legend, false], [color, black], [xlabel, "t"], [ylabel, "A, B, C"], [style, [lines, 4]] )}$

where the plot option \texttt{discrete} informs \texttt{plot2d} that data points rather than a function is submitted as the argument. The resulting graph is similar to Fig. 11.9.

There remains only the question of accuracy. The \texttt{MAXIMA} manuals say only that \texttt{rk} uses a fourth-order Runge-Kutta method\(^\text{46}\) and do not identify option variables that might give the user control over that accuracy. Comparison of the results generated by \texttt{MAXIMA} in the above example with those generated for the same problem by other methods of known accuracy, however, supports the conclusion that the results obtained in the above example are characterized by an absolute accuracy of no less than \(\pm 1.0\) and perhaps of \(\pm 0.1\).

\(^{46}\)See Section 11.6.3 for a description of the method and a brief discussion of its accuracy.
11.16 Exercises

11.16.1 ... using Symbolic Methods

11.1. Find the motion of a driven, damped oscillator satisfying the differential equation

\[ m \frac{d^2 x}{dt^2} + b \frac{dx}{dt} + kx = mf \cos \omega t \]

subject to the general initial conditions

\[ x(0) = x_0 ; \quad \frac{dx}{dt}(0) = v_0 \]

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 (\( m_1 \) and \( m_2 \), 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

\[ m_1 \frac{d^2 x_1}{dt^2} = k(x_2 - x_1) ; \quad m_2 \frac{d^2 x_2}{dt^2} = -k(x_2 - x_1) \]

Let the system be put into motion with arbitrary initial conditions

\[ x_1(0) = x_{10} ; \quad x_2(0) = x_{20} ; \quad \frac{dx_1}{dt}(0) = v_{10} ; \quad \frac{dx_2}{dt}(0) = v_{20} \]

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 transforms of an expression like

\[ \frac{As + Bs^3}{\alpha s^4 + \beta s^2 + \gamma} \]

where \( \alpha, \beta, \gamma \), and \( A, B \) are constants, and 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

\[ s \left( \frac{A + Bs^2}{\alpha s^4 + \beta s^2 + \gamma} \right) = s \left( \frac{R}{s^2 + R'} + \frac{Q}{s^2 + Q'} \right) = \frac{Rs}{s^2 + R'} + \frac{Qs}{s^2 + Q'} \]

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

\[
m_1 \frac{d^2 x_1}{dt^2} = -kx_1 + k(x_2 - x_1) \\
m_2 \frac{d^2 x_2}{dt^2} = -k(x_2 - x_1) + k(x_3 - x_2) \\
m_3 \frac{d^2 x_3}{dt^2} = -k(x_3 - x_2) - kx_3
\]

**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

\[
\frac{d^2 \theta}{dt^2} + \frac{m_2}{m_1 + m_2} \frac{d^2 \phi}{dt^2} + \frac{g}{l} \theta = 0 \\
\frac{d^2 \phi}{dt^2} + \frac{m_1}{m_1 + m_2} \frac{d^2 \theta}{dt^2} + \frac{g}{l} \phi = 0
\]

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
\[ x(0) = x_0 ; \quad \frac{dx}{dt}(0) = 0 \]
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
\[ m \frac{d^2 r}{dt^2} = \frac{qQ}{4\pi \epsilon_0} \frac{r}{r^3} \]
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) \quad \text{Show that the polar components of the equations of motion in the first instance are}
\[ m \left( \frac{d^2 r}{dt^2} - r \left( \frac{d\phi}{dt} \right)^2 \right) = f(r) ; \quad m \left( r \frac{d^2 \phi}{dt^2} + 2 \frac{d\phi}{dt} \frac{dr}{dt} \right) = 0 \]
\[(b) \quad \text{Recognizing that}
\[ \frac{d}{dt} \left( mr^2 \frac{d\phi}{dt} \right) = m \left( r^2 \frac{d^2 \phi}{dt^2} + 2r \frac{dr}{dt} \frac{d\phi}{dt} \right) \]
Show that
\[mr^2 \frac{d\phi}{dt} = \text{constant} = L \quad \implies \quad \frac{d\phi}{dt} = \frac{L}{mr^2}\]
and then that
\[m \frac{d^2r}{dt^2} = f(r) + \frac{L^2}{mr^3}\]
Here, \(L = m(x_0v_{x0} - y_0v_{y0})\) 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 \(d\phi/dt = L/mr^2\)—see Chapter 13—to find \(\phi(t)\).

(c) Translate the initial conditions to polar coordinates, finding that
\[r(0) = \sqrt{x_0^2 + y_0^2}\quad \frac{dr}{dt}(0) = \frac{x_0v_{x0} + y_0v_{y0}}{\sqrt{x_0^2 + y_0^2}}\]
\[\phi(0) = \tan\frac{y_0}{x_0}\quad \frac{d\phi}{dt}(0) = \frac{x_0v_{y0} - y_0v_{x0}}{x_0^2 + y_0^2}\]

(d) Restrict the force to the inverse square gravitational force, finding that
\[m \frac{d^2r}{dt^2} = -\frac{GM}{r^2} + \frac{L^2}{mr^3}\]

(e) Recast the differential equations and initial conditions in dimensionless form, finding that
\[\frac{d^2r}{dt^2} = -\frac{1}{r^2} + \frac{\beta}{r^3} ; \quad \frac{d\phi}{dt} = \pm \sqrt{\beta} \frac{r^2}{r^3}\]
\[r(0) = \frac{r_0}{\ell} = \sqrt{\frac{x_0^2 + y_0^2}{\ell}} ; \quad \phi(0) = \tan\frac{y_0}{x_0}\]
\[\frac{dr}{dt}(0) = \frac{x_0v_{x0} + y_0v_{y0}}{\sqrt{x_0^2 + y_0^2} \sqrt{GM/\ell}} ; \quad \frac{d\phi}{dt}(0) = \frac{x_0v_{y0} - y_0v_{x0}}{(x_0^2 + y_0^2) \sqrt{GM/\ell^3}}\]
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
\[\frac{d^2y}{dx^2} + k^2 y = 0 ; \quad y(0) = y(1) = 0\]
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
\[\frac{d^2y}{dx^2} \bigg|_{x=x_i} \approx \frac{y_{i+1} - 2y_i + y_{i-1}}{\Delta x^2} ; \quad i = 1, 2, 3, \ldots n - 1\]
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

\[
\begin{array}{c}
A \\
\downarrow k_1 \\
B \\
\downarrow k_3 \\
C \\
\downarrow k_2 \\
D
\end{array}
\]

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 \( F = 0 \) to be \( x(0) = 1 \) and \( v(0) = 0 \), produce graphs of \( x \) versus \( t \) 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 \( \bar{t} = \omega t \), 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 \( F = 0 \) 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

\[
\frac{d^2\bar{x}}{dt^2} = \frac{dx}{dt} (1 - \bar{x}^2) - \bar{x}
\]

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\(^\circ\), 45\(^\circ\), 90\(^\circ\), 120\(^\circ\), 150\(^\circ\), 165\(^\circ\), and 178\(^\circ\). 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 \( \bar{t} = \sqrt{g/l} t \) so that the equation becomes \( d^2\theta/d\bar{t}^2 + \sin \theta = 0 \).

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

\[
\frac{d^2\bar{x}}{d\bar{t}^2} = -\frac{\bar{x}}{(\bar{x}^2 + \bar{y}^2)^b} ; \quad \frac{d^2\bar{y}}{d\bar{t}^2} = -\frac{\bar{y}}{(\bar{x}^2 + \bar{y}^2)^b}
\]

[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
b = 1.45, b = 1.55, or anything else you can think of, and write a paragraph or two describing the nature of the changes in some detail. Make sure your solutions are generated to an adequate accuracy to support your conclusions.

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.

11.22. Suppose a particle of charge \( q \) and mass \( m \) is injected into a region of space containing constant, crossed electric and magnetic fields \( \mathbf{E} = E_x \mathbf{i} \) and \( \mathbf{B} = B_z \mathbf{k} \). In vector form, the equation of motion for this particle is

\[
m \frac{d^2 \mathbf{r}}{dt^2} = q \mathbf{E} + q \frac{d\mathbf{r}}{dt} \times \mathbf{B}
\]

Verify the equations of motion

\[
m \frac{d^2 x}{dt^2} = qE_x + qB_z \frac{dy}{dt} ; \quad m \frac{d^2 y}{dt^2} = -qB_z \frac{dx}{dt} ; \quad m \frac{d^2 z}{dt^2} = 0
\]

for the specific fields of this exercise, express them in dimensionless form (note that \( \omega = qB_z/m \) 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 \( qE_x/(m \omega^2 \ell) \), which can alternatively be written as \( (E_z/B_z)/(\omega \ell) \) — the ratio of the velocity \( E_z/B_z \) 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.

11.23. A particle having mass \( m \) and carrying charge \( q \) moves in the \( xy \)-plane while experiencing an electric field given by \( \mathbf{E}(x, y) = -\alpha y \mathbf{j} \), where \( \alpha \) is a constant. Assume that \( \alpha \) and \( q \) are both positive. With \( \mathbf{r} = x \mathbf{i} + y \mathbf{j} \), the vector equation of motion for this particle then is

\[
m \frac{d^2 \mathbf{r}}{dt^2} = -q\alpha y \mathbf{j}
\]

(a) Show that, in component form, the equations of motion for this particle are

\[
\frac{d^2 x}{dt^2} = 0 ; \quad \frac{d^2 y}{dt^2} = -\frac{\alpha q}{m} y = -by
\]

(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.

11.24. 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 \( c = 8.0/3.0 \).
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 \( x \), \( x \) 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. (The method required to produce such a plot
is discussed in Section 3.17.2.)

11.27. The dynamics of the chemical reaction

\[
A + B \leftrightarrow C + D
\]

is governed by the equations

\[
\frac{dA}{dt} = -k_f AB + k_r CD \\
\frac{dB}{dt} = -k_f AB + k_r CD \\
\frac{dC}{dt} = k_f AB - k_r CD \\
\frac{dD}{dt} = k_f AB - k_r CD
\]

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

\[
\frac{dx}{dt} = -k_1 x + k_2 xy \\
\frac{dy}{dt} = k_3 y - k_4 xy
\]

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

\[
m \frac{d^2 x}{dt^2} = -kx + \frac{kax}{\sqrt{x^2 + y^2}} \\
m \frac{d^2 y}{dt^2} = mg - ky + \frac{kay}{\sqrt{x^2 + y^2}}
\]
Then, introducing $\omega_0^2 = k/m$, $\tilde{t} = \omega_0 t$, $\tilde{x} = x/a$, and $\tilde{y} = y/a$, 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

$$\frac{dp}{dt} = f \quad \text{and} \quad \frac{dx}{dt} = v \quad (13.1)$$

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

$$p = \frac{mv}{\sqrt{1 - v^2/c^2}} \quad \text{or} \quad p = mv \quad (13.2)$$
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

\[
m \frac{d^2x}{dt^2} = f(x) \quad \Rightarrow \quad m \frac{dv}{dx} = f(x) \quad \Rightarrow \quad m v \, dv = f(x) \, dx
\]  

(13.5)

Finally, by integrating this last expression from initial values to general values at some other time, we find that

\[
m \int_{v_0}^v v' \, dv' = \frac{1}{2} m v^2 - \frac{1}{2} m v_0^2 = \int_{x_0}^x f(x') \, dx'
\]  

(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

\[
\frac{dx}{dt} = v(x) \quad \Rightarrow \quad dt = \frac{dx}{v(x)} \quad \Rightarrow \quad \int_{t_0}^t \, dt' = t - t_0 = \int_{x_0}^x \frac{dx'}{v(x')}
\]  

(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

\[
\mathbf{r}_{\text{cm}} = \frac{1}{M} \int \sigma(r, \phi) \, r \, dr \, d\phi = \frac{1}{M} \int \sigma(r, \phi) \left( r \cos \phi \hat{i} + r \sin \phi \hat{j} \right) \, r \, dr \, d\phi
\]  

(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

\[
\mathbf{r}_{cm} = \frac{1}{M} \int_0^a \int_0^\pi \frac{M}{2 \pi a^2} \left( r \cos \phi \hat{i} + r \sin \phi \hat{j} \right) r \, dr \, d\phi = \frac{2}{\pi a^2} \int_0^a \int_0^\pi r^2 \left( \cos \phi \hat{i} + \sin \phi \hat{j} \right) r \, dr \, d\phi
\]

(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

\[
\frac{\mathbf{r}_{cm}}{a} = \frac{2}{\pi} \int_0^1 \int_0^\pi \lambda^2 \left( \cos \phi \hat{i} + \sin \phi \hat{j} \right) d\phi d\lambda
\]

(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

\[
I = Mk^2 \quad \Rightarrow \quad k = \sqrt{\frac{I}{M}}
\]

(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

\[
 I = \frac{4M}{\pi a^2} \int_0^a y^2 \sqrt{a^2 - y^2} \, dy = \frac{4}{\pi} Ma^2 \int_0^1 \lambda^2 \sqrt{1 - \lambda^2} \, d\lambda
\]

\[
 \Rightarrow \frac{I}{Ma^2} = \frac{4}{\pi} \int_0^1 \lambda^2 \sqrt{1 - \lambda^2} \, d\lambda
\]

(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

\[
 \int_0^{T/4} \, dt = \frac{T}{4} = \sqrt{\frac{l}{2g} \int_0^{\theta_0} \frac{d\theta}{\sqrt{\cos \theta - \cos \theta_0}}}
\]

(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

\[
\frac{T(k)}{T_0} = 2 \pi \int_0^{\pi/2} \frac{d\phi}{\sqrt{1 - k^2 \sin^2 \phi}} = 2 \pi K(k)
\]  

(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 *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 *Fresnel integrals*, appear in the study of Fresnel diffraction. Together, they define the *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 \( dq \) located at \( \mathbf{r} \) and, in mks units, find the electric field \( \mathbf{E}(\mathbf{r}) \) and the electrostatic potential \( V(\mathbf{r}) \) at the point \( \mathbf{r} \) by evaluating the integrals

\[
\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi \varepsilon_0} \int \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3} \, dq' \quad \text{and} \quad V(\mathbf{r}) = \frac{1}{4\pi \varepsilon_0} \int \frac{dq'}{|\mathbf{r} - \mathbf{r}'|} \quad (13.28)
\]

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 \( \mathbf{B}(\mathbf{r}) \) and the magnetic vector potential \( \mathbf{A}(\mathbf{r}) \) at the point \( \mathbf{r} \) by evaluating the integrals

\[
\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int \frac{I'}{|\mathbf{r} - \mathbf{r}'|^3} \, \mathbf{d}\mathbf{r}' \times (\mathbf{r} - \mathbf{r}') \quad \text{and} \quad \mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int \frac{I'}{|\mathbf{r} - \mathbf{r}'|} \, d\mathbf{r}' \quad (13.29)
\]

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 \( \mathbf{r} = r \cos \phi \, \mathbf{i} + r \sin \phi \, \mathbf{j} + z \, \mathbf{k} \), produced by *one* such ring positioned in the \( xy \) plane with its center at the origin. Let \( \mathbf{r}' = a \cos \phi' \, \mathbf{i} + a \sin \phi' \, \mathbf{j} \) locate an element on that (single) ring. In this notation,

\[
\mathbf{r} - \mathbf{r}' = (a \cos \phi - a \cos \phi') \, \mathbf{i} + (a \sin \phi - a \sin \phi') \, \mathbf{j} + z \, \mathbf{k}
\]

so

\[
|r - r'| = \left[ (r \cos \phi - a \cos \phi')^2 + (r \sin \phi - a \sin \phi')^2 + z^2 \right]^{1/2}
\]

\[
= \left[ r^2 + a^2 - 2ar \cos(\phi' - \phi) + z^2 \right]^{1/2}
\]

(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

\[
V_{\text{one}}(r, \phi, z) = \frac{1}{4\pi \varepsilon_0} \int_0^{2\pi} \frac{\lambda a \, d\phi'}{|r^2 + a^2 - 2ar \cos(\phi' - \phi) + z^2|^{1/2}}
\]

(13.32)
Figure 13.3: The potential produced by two charged rings. Part (a) shows the geometry of ultimate interest. Part (b) shows the simple situation with one ring used as a stepping stone.

The entire integral assumes a simpler appearance, however, on the variable \( \alpha = \phi' - \phi \), becoming

\[
V_{\text{one}}(r, \phi, z) = \frac{\lambda a}{4\pi \epsilon_0} \int_0^{2\pi} \frac{d\alpha}{[r^2 + a^2 - 2ar \cos \alpha + z^2]^{1/2}}
\] (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

\[
V_{\text{midplane}}(r) = V_{\text{one}}(r, \phi, -a) + V_{\text{one}}(r, \phi, a) = \frac{2\lambda a}{4\pi \epsilon_0 a} \int_0^{2\pi} \frac{d\alpha}{[r^2 + 2a^2 - 2ar \cos \alpha]^{1/2}}
\] (13.34)

Finally, introducing the variable \( s \) defined by \( r = sa \) to express the radial coordinate in dimensionless terms, we find the expression

\[
V_{\text{two}}(s) = V_{\text{midplane}}(sa) = \frac{2\lambda}{4\pi \epsilon_0 a} \int_0^{2\pi} \frac{d\alpha}{[s^2 - 2s \cos \alpha + 2]^{1/2}}
\] (13.35)

or, even better, the expression

\[
V(s) = \frac{\sqrt{2}}{2\pi} \int_0^{2\pi} \frac{d\alpha}{[s^2 - 2s \cos \alpha + 2]^{1/2}}
\] (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

\[
P(x_1 < x < x_2) = \int_{x_1}^{x_2} |\psi(x)|^2 \, dx
\]

(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

\[
\psi(x) = \left( \frac{m\omega}{\pi \hbar} \right)^{1/4} e^{-y^2/2}
\]

(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 (\( 1/2 \hbar \omega \) for the ground state) is equal to the potential energy \( 1/2 kx_{\text{turn}}^2 \), the turning point of this oscillator is given by

\[
\frac{1}{2} \hbar \omega = \frac{1}{2} kx_{\text{turn}}^2 \implies x_{\text{turn}} = \sqrt{\frac{\hbar \omega}{k}} = a
\]

(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

\[
P(\{|x| > |x_{\text{turn}}|\}) = \int_{-\infty}^{-a} |\psi(x)|^2 \, dx + \int_{a}^{\infty} |\psi(x)|^2 \, dx = 1 - \int_{-a}^{a} |\psi(x)|^2 \, dx
\]

(13.40)
Finally, after substituting the wave function and recasting the entire integral as an integral on the variable \( y \), we find that

\[
P(\{|x| > |x_{\text{turn}}|\}) = 1 - \frac{1}{\sqrt{\pi}} \int_{-1}^{1} e^{-y^2} \, dy = 1 - \frac{2}{\sqrt{\pi}} \int_{0}^{1} e^{-y^2} \, dy = 1 - \text{erf}(1)
\]

(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

\[
\int_{a}^{b} \phi_i(x) \phi_j(x) \, w(x) \, dx = N_j \delta_{ij}
\]

(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

\[
f(x) = \sum_{n} a_n \phi_n(x)
\]

(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

\[
\int_a^b w(x) \phi_j(x) f(x) \, dx = \sum_n a_n \int_a^b w(x) \phi_j(x) \phi_n(x) \, dx = \sum_n a_n N_n \delta_{nj} = N_j a_j \tag{13.44}
\]

We conclude that, if we know either of \( f(x) \) or \( a_n \), we can determine the other, i.e., that

\[
f(x) = \sum_n a_n \phi_n(x) \quad \Leftrightarrow \quad a_n = \frac{1}{N_n} \int_a^b w(x) \phi_n(x) f(x) \, dx \tag{13.45}
\]

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 \(^1\) is

\[
\left\{ 1, \sin \frac{n \pi x}{l}, \cos \frac{n \pi x}{l} \; ; \; n, m = 1, 2, 3, \ldots \right\} \tag{13.46}
\]

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

\[
\int_{-l}^l 1 \times 1 \, dx = 2l \quad ; \quad \int_{-l}^l 1 \times \sin \frac{n \pi x}{l} \, dx = 0 \quad ; \quad \int_{-l}^l 1 \times \cos \frac{n \pi x}{l} \, dx = 0 \tag{13.47}
\]

\[
\int_{-l}^l \sin \frac{n \pi x}{l} \, \sin \frac{m \pi x}{l} \, dx = l \delta_{nm} \quad ; \quad \int_{-l}^l \sin \frac{n \pi x}{l} \, \cos \frac{m \pi x}{l} \, dx = 0 \tag{13.48}
\]

\[
\int_{-l}^l \cos \frac{n \pi x}{l} \, \cos \frac{m \pi x}{l} \, dx = l \delta_{nm} \quad ; \quad \int_{-l}^l \cos \frac{n \pi x}{l} \, \sin \frac{m \pi x}{l} \, dx = 0 \tag{13.49}
\]

The expansion\(^2\)

\[
f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left( a_n \cos \frac{n \pi x}{l} + b_n \sin \frac{n \pi x}{l} \right) \tag{13.50}
\]

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

\[
a_n = \frac{1}{l} \int_{-l}^l f(x) \cos \frac{n \pi x}{l} \, dx \quad ; \quad b_n = \frac{1}{l} \int_{-l}^l f(x) \sin \frac{n \pi x}{l} \, dx \tag{13.51}
\]

### 13.2 Evaluating Integrals Symbolically with MAXIMA

The MAXIMA command for symbolic integration is `integrate`. It takes either two or four arguments, the first of which supplies the expression to be integrated, the second of which identifies the variable of integration, and—when present—the third and fourth of which specify the lower and upper limits, respectively. Thus, for example, the statement\(^3\)

\[
\text{integrate}( \sin(kx), x );
\]

\(^1\)To put this set into the form of the previous paragraph, we would identify \( \phi_1(x) = 1, \phi_2(x) = \sin(\pi x/l), \phi_3(x) = \cos(\pi x/l), \phi_4(x) = \sin(2\pi x/l), \ldots \)

\(^2\)The first coefficient is written \( a_0/2 \) rather than \( a_0 \) to simplify the integrals determining the coefficients. With the choice we have made, the integral for \( a_0 \) turns out to be obtainable by setting \( n = 0 \) in the expression for \( a_n \); we do not need a third—and different—expression for that one coefficient.

\(^3\)Remember that, in some versions of MAXIMA, the terminating semicolon is unnecessary.
evaluates the indefinite integral $\int \sin(kx) \, dx$ while the statement

```
integrate( \sin(k*x), x, 0, %pi);
```

evaluates the definite integral $\int_0^\pi \sin(kx) \, dx$. If MAXIMA 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. MAXIMA also evaluates improper integrals, recognizing the symbols $\inf$ for $+\infty$ and $\minf$ 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.

In this section, we illustrate the use of the command `integrate` to evaluate some of the integrals deduced in Section 13.1. Beyond integration per se, we also show how, in many cases, MAXIMA can be used to set up the integral as well. To abbreviate the presentation of MAXIMA dialogs, we make liberal use of terminating dollar signs to suppress intermediate output. The reader is therefore urged to duplicate the dialogs in an actual session with MAXIMA, either replacing the dollar signs with semicolons or eliminating them (depending on the version of MAXIMA in use).

### 13.2.1 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 constant force $f$. Then, the basic relationships from which we would determine $x(t)$ and $v(t)$ are

\[
p(t) = \int_0^t f \, dt', \quad x(t) = \int_0^t v(t') \, dt', \quad \text{and} \quad p(t) = \frac{mv(t)}{\sqrt{1 - v(t)^2/c^2}}\tag{13.52}
\]

The following “conversation” with MAXIMA will find the desired quantities and explore a few of their properties:

\[
\text{(i1) } \text{assume( c>0, f>0, m>0, t>0 )$ \quad \text{Declare c, f, m and t to be positive.}}
\]

\[
\text{(i2) } \text{p : integrate(f, tp, 0, t); \quad \text{Integrate (constant) f to find p.}}
\]

\[
\text{(i3) } \text{p = m * v / sqrt(1 - v^2/c^2); \quad \text{Relate momentum to velocity.}}
\]

\[
\text{(i4) } \text{ft; \quad \text{Prepare to solve for } v.}
\]

\[
\text{(i5) } \text{ft = m * v / sqrt(1 - v^2/c^2); \quad \text{Prepare to solve for } v.}
\]

\[
\text{(i6) } \text{v : part( ft, 2, 2 ); \quad \text{Extract second solution. (Physically, } v \text{ and } f \text{ must have the same sign.)}
\]

\[
\text{(i7) } \text{limit( v, t, inf ); \quad \text{Find limit of } v \text{ as } t \to \infty.}
\]

\[
\text{(i8) } \text{vf = subst( tp, t, v ); \quad \text{Substitute } tp \text{ for } t \text{ in } v.}
\]

\[
\text{(i9) } \text{x : integrate( vf, tp, 0, t ); \quad \text{Integrate } vf \text{ to find position.}}
\]

\[
\text{(i10) } \text{cf = f / sqrt(f^2/t^2 + c^2 m^2); \quad \text{Integrate } vf \text{ to find position.}}
\]

\[\text{MAXIMA statements are shown on the left; comments describing the statements are shown on the right.}\]
13.2. EVALUATING INTEGRALS SYMBOLICALLY WITH MAXIMA

Find limit of \( v \) for small \( t \).

\[
\frac{ft}{m} - \frac{f^3 t^3}{2c^2 m^3} + \ldots
\]

Find limit of \( x \) for small \( t \).

\[
\frac{ft^2}{2m} - \frac{f^3 t^4}{8c^2 m^3} + \ldots
\]

Reassuringly, the velocity approaches \( c \) as \( t \to \infty \) and, with the acceleration \( a \) identified as \( f/m \), \( v \) and \( x \) for small \( t \) (classical limit) are \( at \) and \( \frac{1}{2}at^2 \) respectively.

13.2.2 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

\[
\mathbf{rr} : [ r \cos \phi, r \sin \phi, 0];
\]

\[
\mathbf{sigma} : \frac{m}{\left(\frac{\pi}{2} \cdot a^2 / 2\right)};
\]

\[
\mathbf{r_cm} : \int \int (\mathbf{sigma} \cdot \mathbf{rr} \cdot r, \phi, 0, \%pi, r, 0, a) / m;
\]

\[
\mathbf{float}(\%);
\]

Since only the \( y \) component of \( \mathbf{r_cm} \) 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 MAXIMA has at \( \%i3 \) effected an item-by-item integration of the elements of a list with a single statement.

Note also that, in line \( \%i3 \), we have nested one \texttt{integrate} command within another. 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.

13.2.3 Moment of Inertia; Radius of Gyration

The evaluation by MAXIMA of the moment of inertia of a semicircular plate as given by Eq. (13.16) is quick and straightforward. We need only the statements

\[
\mathbf{assume}( a>0, m>0 );
\]

\[
\mathbf{integrate}(\lambda^2 \cdot \sqrt{1-\lambda^2}, \lambda, 0, 1);
\]

\[
\mathbf{I} : \frac{ma^2}{4};
\]

\[
\mathbf{k} : \sqrt{\mathbf{I}/m};
\]

\[
\frac{a}{2}
\]

to evaluate the necessary integral and find both the moment of inertia \( I = ma^2/4 \) and the radius of gyration \( k = a/2 \).
13.2.4 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

$$V(x, y, z) = \frac{\lambda}{4\pi\epsilon_0} \int_{-a}^{+a} \frac{dz'}{|r - r'|} \quad (13.53)$$

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 statements

$(%i1) \ rr : [r*cos(phi), r*sin(phi), z]$
$(%i2) \ rp : [0, 0, zp]$
$(%i3) \ sep : rr - rp$
$(%i4) \ sep . sep$
$(%i5) \ trigsimp(%)$
$(%i6) \ denom : sqrt(%)$
$(%i7) \ cnst : 1/(4*%pi*epsilon[0])$
$(%i8) \ integ : lambda * cnst / denom;$

Assign $r$ in cylindrical coordinates.
Assign $r'$.
Evaluate $r - r'$.
Evaluate $|r - r'|^2$.
Recognize $\cos^2 \phi + \sin^2 \phi = 1$.
Bind $|r - r'|$ to denom.
Bind $1/4\pi\epsilon_0$ to cnst.
Set integrand. We use $\lambda$ in the output even though MAXIMA does not translate lambda into $\lambda$.

$(%i9) \ assume( r>0, z>0, a>0 )$
$(%i10) \ v : integrate( integ, zp, -a, a);$  

Declare $r$, $z$, and $a$ to be positive.
Find $V(r, z)$, with $r = \sqrt{x^2 + y^2}$.

$$\lambda \frac{\text{asinh}[(z + a)/r] - \text{asinh}[(z - a)/r]}{4\pi\epsilon_0}$$

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
13.2. EVALUATING INTEGRALS SYMBOLICALLY WITH MAXIMA

\[(\%i11) \quad v_0 : v, z = 0; \quad \text{Evaluate potential in } xy \text{ plane.} \]

\[(\%o11) \quad \frac{\lambda}{2\pi\epsilon_0} \text{asinh} \left( \frac{a}{r} \right) \]

\[(\%i12) \quad \text{subst( s*r, a, v0 )} \quad \text{Let } a/r \to s. \]

\[(\%i13) \quad \text{taylor( , s, 0, 2 )} \quad \text{Expand near } s = 0 \ (a/r \to 0). \]

\[(\%i14) \quad \text{subst( a/r, s, )} \quad \text{Return to original variables.} \]

\[(\%i15) \quad \text{subst( q/(2*a), lambda, )} \quad \text{Express in terms of total charge.} \]

\[(\%o15) \quad \frac{q}{4\pi\epsilon_0 r} \]

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 would expand the potential with the statement

\[(\%i16) \quad \text{logcontract( taylor( v0, r, 0, 1 ) );} \quad \text{Expand potential at } z = 0 \text{ for small } r, \text{ keeping only first non-zero term.} \]

\[(\%o16) \quad \frac{\lambda}{2\pi\epsilon_0} \log \left( \frac{2a}{r} \right) \]

or, equivalently

\[ -\frac{\lambda}{2\pi\epsilon_0} \log \frac{r}{a} + \frac{\lambda \log 2}{2\pi\epsilon_0} \]

This 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.

13.2.5 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 \( \mathbf{B}(\mathbf{r}) \) 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 \( \mathbf{r} = sa \mathbf{k}, \) 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\(^5\)

5The file $HEAD/maxima/crossdot.mac was described in Section 6.9 and exists in the public directory structure. (The symbol $HEAD is defined in the Local Guide.) It defines the two functions cross(v1, v2) and dot(v1, v2) for evaluating the cross and dot products of two three-component vectors.
(%i12) load("$HEAD/maxima/crossdot.mac")$  
Define cross and dot.
(\%i13) assume( a > 0 )$  
Declare a positive.
(\%i14) r : [0, 0, s * a]$
Assign \( \mathbf{r} \).
(\%i15) rp : [a*cos(phi), a*sin(phi), 0]$  
Assign \( \mathbf{r}' \).
(\%i16) sep : r - rp$
Evaluate \( \mathbf{r} - \mathbf{r}' \).  
(\%i17) dot( sep, sep )$
Evaluate \( |\mathbf{r} - \mathbf{r}'|^2 \).  
(\%i18) trigsimp(%)$
Recognize \( \cos^2 \phi + \sin^2 \phi = 1 \).
(\%i19) mag2 : factor(%)$  
\( \text{mag}^2 = |\mathbf{r} - \mathbf{r}'|^2 \).
(\%i10) cross( drp, sep )$
Evaluate \( \mathbf{d}r' \times (\mathbf{r} - \mathbf{r}') \).  
(\%i11) num : trigsimp(%);$  
Recognize \( \cos^2 \phi + \sin^2 \phi = 1 \).

Finally, we bring various factors together to construct the integrand, evaluate the integral, and recast the result in a dimensionless form with the statements

(\%i12) integrand : mu[0] * i * num / 4 / %pi / mag2^(3/2);  
Integrate around the loop.
(\%i13) b1 : integrate(%,phi,0,2*%pi);  
Integrate around the loop.
(\%i14) b1z : part(%, 3)$  
Extract the z component.
(\%i15) b1c : b1z, s = 0$  
Find field at center.
(\%i16) b1z : b1z / b1c;  
Normalize field of single loop.
(\%i17) bpair : subst(s-c, s, b1z)$  
Find field of two loops, one at \( s = c \) and the other at \( s = -c \).
(\%i18) bpairc : bpair, s = 0$  
Find field at center.
(\%i19) bpair : bpair / bpairc;  
Normalize field of two loops.
(\%o19) \( \frac{(c^2 + 1)^{3/2}}{2} \left( \frac{1}{|(s + c)^2 + 1|^{3/2}} + \frac{1}{|(s - c)^2 + 1|^{3/2}} \right) \).

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.
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 \( c = 0.25 \). Here, the default labels on the \( x \) and \( y \) axes stand for \( z/a \) and \( B_{\text{pair}}/B_{\text{pairc}} \), respectively.

\[
\text{(\%i20)} \quad c100 : \text{bpair, c}=1.0$
\]
\[
\text{(\%i21)} \quad c050 : \text{bpair, c}=0.5$
\]
\[
\text{(\%i22)} \quad c025 : \text{bpair, c}=0.25$
\]
\[
\text{(\%i23)} \quad \text{plot2d( } [\text{c100, c050, c025 }], [s, 0.0, 3.0], [y, 0.0, 2.0],$
\]
\[
\quad \quad \quad \quad \text{[style, [lines, 4]], [color, black], [legend, false],}$
\]
\[
\quad \quad \quad \quad \text{[xlabel, \"s = z/a\"], [ylabel, \"B(z) / B(0)\"] );$
\]

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

\[
\text{(\%i24)} \quad \text{taylor(bpair, s, 0, 4)}$
\]
\[
\text{(\%o24)} \quad 1 + \frac{(12c^2 - 3)s^2}{2c^4 + 4c^2 + 2} + \frac{(120c^4 - 180c^2 + 15)s^4}{8c^6 + 32c^4 + 48c^4 + 32c^2 + 8} + \ldots$
\]

Clearly there is a particular value of \( c \) 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
Extract first factor of numerator of second term of Taylor series. Note that the numbers in the argument of `part` will depend on how `%o24` is arranged by your system.

Find c to make numerator zero.

The two solutions are 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 how the field varies with \( s \) (when \( s \) is small) by using the statements

\[
(\%i27) \quad \text{Evaluate field for Helmholtz separation.}
\]

\[
(\%o27) \quad \frac{-144 s^4 - 125}{125}
\]

\[
(\%i28) \quad \text{Cast result in floating form.}
\]

\[
(\%o28) \quad 1.0 - 1.152 s^4
\]

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

\[
(\%i29) \quad \text{Find } s \text{ when field is down 1%.}
\]

\[
(\%o29) \quad [0.3052367917903921, ...]
\]

\[
(\%i30) \quad \text{Find } s \text{ when field is down 5%}
\]

\[
(\%o30) \quad [0.4564354645876384, ...]
\]

MAXIMA, of course, returns a list containing all four of the solutions to the quartic equation. We have displayed the only solution that is both real and positive.

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 `bpair` but restrict it to the Helmholtz case by setting \( c = 1/2 \) with the statement

\[
(\%i31) \quad \text{bhelm : bpair, } c = 1/2$
\]

Then, we seek the values of \( s \) at which this expression assumes the value 0.99 (1% fall off) or 0.95 (5% fall 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 `solve` is not up to the task. Thus, we draw (prematurely—see Section 14.11) on MAXIMA’s command `find_root`, which uses a binary search to find a root in a specified range. The above results provide us with sensible initial guesses, so we invoke the statements\(^{6}\)

\[^{6}\text{In `find_root`, the first argument is the expression whose root is sought, the second is the variable to be adjusted in finding the root, and the third and fourth provide the range within which a root is sought. By default, `find_root` strives to find the root to absolute and relative tolerances of 0.0 and 0.0.}\]
13.2. EVALUATING INTEGRALS SYMBOLICALLY WITH MAXIMA

(%i32) find_root(bhelm-0.99, s, 0.3, 0.35);  
Find s when field is down 1%.
(%o32) 0.313746025198168

(%i33) find_root(bhelm-0.95, s, 0.4, 0.5)  
Find s when field is down 5%.
(%o33) 0.48846550123516077

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

(%i34) float( ev(bhelm, s=0.5) );  
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 \( s^4 = (\pi/a)^4 \) As the results at \(%o32-%o34\) 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.

13.2.6 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. MAXIMA 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

(%i11) integrand : 1/sqrt(1-k^2*sin(phi)^2);
(%o11) \frac{1}{\sqrt{1 - k^2 \sin^2 \phi}}

(%i12) taylor( %, k, 0, 4 );
(%o12) /T/ 1 + \left( \sin^2 \phi \right) k^2 + \left( 3 \sin^4 \phi \right) k^4 + \ldots

Then we insert the pre-multiplying factor \( 2/\pi \) and integrate the series on \( \phi \) with the statement

(%i13) expand( 2*integrate( %, phi, 0, %pi/2 ) / %pi );
(%o13) 1 + \frac{k^2}{4} + \frac{9 k^4}{64} + \ldots

Finally, we substitute \( \sin(\theta_0/2) \) for \( k \) with the statement

(%i14) subst(sin(theta[0]/2),k,%);
(%o14) 1 + \frac{1}{4} \sin^2(\theta_0/2) + \frac{9}{64} \sin^4(\theta_0/2) + \ldots

and expand this result as a power series in \( \theta_0 \)
CHAPTER 13. EVALUATING INTEGRALS

(%i5) taylor(%, theta[0], 0, 4);
(%o5) /T/ 1 + 16 θ^2 + 11/3072 θ^4 + ...

Note that θ₀ must here be expressed in radians.

To see how significantly the period departs from the small-amplitude value when θ₀ is not quite zero, let us evaluate this expression for θ₀ = 5°, 10°, 15°, and 20°. The MAXIMA statements

(%i6) [5, 10, 15, 20] $
(%i7) ampl : (%/180)*%pi $
(%i8) per1 : float(1+ampl^2/16)
(%o8) [1.000476, 1.001904, 1.004284, 1.007615]
(%i9) per2 : per1 + float(11*ampl^4/3072);
(%o9) [1.000476, 1.001907, 1.004301, 1.007669]

Set angles in degrees.
Convert angles to radians.
Calculate first correction. Values are rounded to six digits after the decimal point.
Add second correction.

Evidently, at amplitudes of 5°, 10°, 15°, and 20°, the period of the pendulum is approximately 0.05%, .2%, .4%, and .8% larger than the standard small-amplitude approximation given by the expression $2\pi \sqrt{\frac{l}{g}}$.

13.2.7 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

$$a_0 = \frac{1}{l} \int_0^l \sin \frac{\pi x}{l} \, dx ; \quad a_n = \frac{1}{l} \int_0^l \sin \frac{\pi x}{l} \cos \frac{n\pi x}{l} \, dx ; \quad b_n = \frac{1}{l} \int_0^l \sin \frac{\pi x}{l} \sin \frac{n\pi x}{l} \, dx$$

(13.55)

where $n = 1, 2, 3, \ldots$. We find these coefficients quickly by submitting to MAXIMA the statements

(%i11) declare( n, integer )$
(%i12) assume( l>0 )$
(%i13) integrate(sin(%pi*x/l),x,0,l)/l;
(%i14) integrate(sin(%pi*x/l)*cos(n*%pi*x/l),x,0,l)/l;
(%i15) factor(%); 2/π
(%i16) integrate(sin(%pi*x/l)*cos(%pi*x/l),x,0,l)/l;
(%i17) integrate(sin(%pi*x/l)*sin(n*%pi*x/l),x,0,l)/l;
(%i18) integrate(sin(%pi*x/l)*sin(%pi*x/l),x,0,l)/l;
(%i18) 1/2

Force n to be an integer.
Declare l > 0.
Find a₀. Note the distinction between l (el) and 1 (one).
Here and in the next several statements, the character is el.
Find aₙ, n > 0.
Find aₙ, n = 1.
Find bₙ, n ≠ 1.
Find bₙ, n = 1.
Here, the first statement limits the nature of \( n \) so that MAXIMA can automatically replace expressions like \( \sin n\pi \) and \( \cos n\pi \) with explicit evaluations, the second statement assures MAXIMA that \( l \) is positive, and the remaining statements evaluate the desired integrals. Thus, we find from the output on lines \( \%o3 \), \( \%o5 \), \( \%o6 \), \( \%o7 \), and \( \%o8 \) that

\[
a_0 = \frac{2}{\pi} ; \quad a_1 = 0 ; \quad a_n = \frac{-1 + (-1)^n}{n(n^2 - 1)}, n > 1 ; \quad b_1 = \frac{1}{2} ; \quad b_n = 0, n > 1 \quad (13.56)
\]

With these values, we assemble the Fourier series in accordance with Eq. (13.50) to find that

\[
f(x) = \frac{1}{\pi} - \sum_{n=2}^{\infty} \frac{1 + (-1)^n}{n(n^2 - 1)} \cos \frac{n\pi x}{l} + \frac{1}{2} \sin \frac{\pi x}{l}
\]

\[
= \frac{1}{\pi} + \frac{1}{2} \sin \frac{\pi x}{l} - \frac{2}{\pi} \sum_{m=1}^{\infty} \frac{1}{4m^2 - 1} \cos \frac{2m\pi x}{l}
\]

\[
= \frac{1}{\pi} + \frac{1}{2} \sin \frac{\pi x}{l} - \frac{2}{\pi} \left( \frac{1}{3} \cos \frac{2\pi x}{l} + \frac{1}{15} \cos \frac{4\pi x}{l} + \frac{1}{35} \cos \frac{6\pi x}{l} + \cdots \right) \quad (13.57)
\]

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

\[
f_n(x) = \frac{1}{\pi} + \frac{1}{2} \sin \frac{\pi x}{l} - \frac{2}{\pi} \sum_{m=1}^{n} \frac{1}{4m^2 - 1} \cos \frac{2m\pi x}{l}
\]

which approaches \( f(x) \) as \( n \to \infty \). Introducing the variable \( \pi = x/l \) (and then dropping the overbar), we then generate graphs for various values of \( n \) with the statements listed in Table 13.1. Here, we force \( m \) to be an integer, recast the general coefficient in \( \%o4 \) by replacing \( n \) with \( 2m \), construct a MAXIMA function defining the truncated series, and plot the truncated series for each of four different truncation points. 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 formulæ—for numerical integration. In this section, we describe several of these formulæ.

#### 13.5.1 Newton-Cotes Quadrature

One family of quadrature formulæ can be deduced by starting with the recognition that the definite integral

\[
A = \int_{a}^{b} f(x) \, dx \quad (13.59)
\]

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
Table 13.1: Graphing several Fourier series with MAXIMA.

(%i9) declare( m, integer )$
(%i10) factor( subst( 2*m, n, %o5 ) );
(%o10) \frac{2}{\pi(2m-1)(2m+1)}$
(%i11) a(m) := -2/(%pi*(2*m-1)*(2*m+1))$
(%i12) f(x,n) := 1/%pi + sin(%pi*x)/2 + sum( a(m)*cos(2*m*%pi*x),m,1,n)$(%i13) plot2d( f(x,1), [x, -3.0, 3.0], [y, -0.5, 1.5],
          [xlabel, "x"], [ylabel, "f(x,1)"], [style, [lines, 4]],
          [legend,false], [color, black] )$
(%i14) plot2d( f(x,2), [x, -3.0, 3.0], [y, -0.5, 1.5],
          [xlabel, "x"], [ylabel, "f(x,2)"], [style, [lines, 4]],
          [legend,false], [color, black] )$
(%i15) plot2d( f(x,4), [x, -3.0, 3.0],
          [y, -0.5, 1.5],
          [xlabel, "x"], [ylabel, "f(x,4)"], [style, [lines, 4]],
          [legend,false], [color, black] )$
(%i16) plot2d( f(x,8), [x, -3.0, 3.0], [y, -0.5, 1.5],
          [xlabel, "x"], [ylabel, "f(x,8)"], [style, [lines, 4]],
          [legend,false], [color, black] )$

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. These graphs were created with MAXIMA.
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

\[
\int_a^b f(x) \, dx \approx f_0 \Delta x + f_1 \Delta x + \cdots + f_i \Delta x + \cdots + f_{N-1} \Delta x
\]

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**

\[
\int_a^b f(x) \, dx \approx M_N = (f_{1/2} + f_{3/2} + f_{5/2} + \cdots + f_{N-3/2} + f_{N-1/2}) \Delta x
\]

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**

\[
\int_a^b f(x) \, dx \approx T_N = \left( \frac{1}{2} f_0 + f_1 \right) \Delta x + \left( \frac{1}{2} f_1 + f_2 \right) \Delta x + \cdots + \left( \frac{1}{2} f_{N-1} + f_N \right) \Delta x
\]
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,

\[
\int_a^b f(x) \, dx \approx S_N = \frac{1}{3} \left( f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots + 2f_{N-2} + 4f_{N-1} + f_N \right) \Delta x
\]

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

\[
T_2 = T_{2^1} = \left[ \frac{f(a)}{2} + f(x_1) + \frac{f(b)}{2} \right] \frac{(b - a)}{2} = \frac{1}{2} T_1 + f(x_1) \frac{(b - a)}{2}
\]

\[
= \frac{1}{2} T_1 + f \left( a + \frac{b - a}{2} \right) \frac{(b - a)}{2} \quad (13.65)
\]

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

\[
T_1 = T_{2^2} = \left[ \frac{f(a)}{2} + f(x_2) + f(x_1) + f(x_3) + \frac{f(b)}{2} \right] \frac{(b - a)}{4}
\]

\[
= \frac{1}{2} T_2 + \left[ f(x_2) + f(x_3) \right] \frac{(b - a)}{4}
\]

\[
= \frac{1}{2} T_2 + \left[ f \left( a + \frac{b - a}{4} \right) + f \left( a + 3 \frac{b - a}{4} \right) \right] \frac{(b - a)}{4} \quad (13.66)
\]
Yet again, introducing
\[ x_4 = a + \frac{1}{8} (b - a) \quad x_5 = a + \frac{3}{8} (b - a) \]
\[ x_6 = a + \frac{5}{8} (b - a) \quad x_7 = a + \frac{7}{8} (b - a) \] (13.67)
and continuing one more step, we find that
\[
T_8 = T_{2^3} = \frac{1}{2} T_4 + \left[ f(x_4) + f(x_5) + f(x_6) + f(x_7) \right] \frac{(b - a)}{8}
\]
\[
= \frac{1}{2} T_4 + \left[ f \left( a + \frac{b - a}{8} \right) + f \left( a + \frac{3b - a}{8} \right) + f \left( a + \frac{5b - a}{8} \right) + f \left( a + \frac{7b - a}{8} \right) \right] \frac{(b - a)}{8}
\]
(13.68)
and, in general, with
\[ x_{2N+j} = a + \frac{(2j + 1)(b - a)}{2^{N+1}} \quad 0 \leq j < 2^{N+1} \] (13.69)
that
\[
T_{2^{N+1}} = \frac{1}{2} T_{2^N} + \left[ f \left( a + \frac{b - a}{2^{N+1}} \right) + f \left( a + \frac{3b - a}{2^{N+1}} \right) + \ldots \right.
\]
\[
+ \left. f \left( a + (2^N - 1) \frac{b - a}{2^{N+1}} \right) \right] \frac{(b - a)}{2^{N+1}}
\]
(13.70)
Evidently, we can step from any evaluation by the trapezoidal rule to an evaluation by the trapezoidal rule with twice as many divisions without recalculating anything that we have already calculated! We shall refer to this embellishment as the recursive trapezoidal rule.

A second strategy for making algorithms more efficient involves what is called Richardson extrapolation. Using the trapezoidal rule of Eq. (13.62) and labeling the interpolation points as in Fig. 13.9, we find, for example, that
\[
T_4 = \frac{1}{2} f(a) + f(x_2) + f(x_1) + f(x_3) + \frac{1}{2} f(b) \frac{b - a}{4}
\]
\[
= \left[ f(a) + 2f(x_2) + 2f(x_1) + 2f(x_3) + f(b) \right] \frac{b - a}{8}
\]
(13.71)
and that
\[
T_8 = \frac{1}{2} f(a) + f(x_4) + f(x_2) + f(x_5)
\]
\[
+ f(x_1) + f(x_6) + f(x_3) + f(x_7) + \frac{1}{2} f(b) \frac{b - a}{8}
\]
(13.72)
Note, in particular, the combination
\[
\frac{4T_8 - T_4}{3} = \left[ f(a) + 4f(x_4) + 2f(x_2) + 4f(x_5) + 2f(x_1) \right.
\]
\[
+ 4f(x_6) + 2f(x_3) + 4f(x_7) + f(b) \left( \frac{b - a}{8 \times 3} \right)
\]
(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
\[
\frac{4T_8 - T_4}{3} = \left[ f(a) + 4f(x_4) + 2f(x_2) + 4f(x_5) + 2f(x_1) \right.
\]
\[
+ 4f(x_6) + 2f(x_3) + 4f(x_7) + f(b) \left( \frac{\Delta x}{3} \right)
\]
(13.74)
which—

mirabile dictu—

we recognize as Simpson’s rule for evaluating the integral with 8 divisions of the interval! We conclude that

\[ S_8 = \frac{4T_8 - T_1}{3} \]  

(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 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 Romberg integration. We do actual numerical integration 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 four values \( T_n, T_{2n}, T_{4n}, \) and \( T_{8n} \). We could then use Eq. (13.76) to generate the 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. Truncation errors arise because the integral has been approximated by a finite sum; 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 \( 10^4 \). 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.\(^7\) 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).\(^8\)

\(^7\)We shall make this criterion a bit more explicit in the next paragraphs.

\(^8\)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. For the midpoint formula of Eq. (13.61), for example,

\[
\left| \int_a^b f(x) \, dx - M_N \right| = \frac{(b-a)^3}{24N^2} \left| \frac{d^2 f}{dx^2} \right|_{x=\xi}
\]

(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

\[
\left| \int_a^b f(x) \, dx - T_N \right| = \frac{(b-a)^3}{12N^2} \left| \frac{d^2 f}{dx^2} \right|_{x=\xi}
\]

(13.78)

and

\[
\left| \int_a^b f(x) \, dx - S_N \right| = \frac{(b-a)^5}{180N^4} \left| \frac{d^4 f}{dx^4} \right|_{x=\xi}
\]

(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 formulae.

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 formulae 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 \( d^i f/dx^i \) in the interval \( a < x < b \), then the above expressions support the inequalities

\[
\left| \int_a^b f(x) \, dx - M_N \right| \leq \frac{(b-a)^3}{24N^2} D_{\text{max}}(2)
\]

(13.80)

\[
\left| \int_a^b f(x) \, dx - T_N \right| \leq \frac{(b-a)^3}{12N^2} D_{\text{max}}(2)
\]

(13.81)

\[
\left| \int_a^b f(x) \, dx - S_N \right| \leq \frac{(b-a)^5}{180N^4} D_{\text{max}}(4)
\]

(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 $S_2$, $S_4$, $S_8$, ..., 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

$$
\int_a^b g(t) \, dt = \sum_{i=1}^m \int_{t_i}^{t_{i+1}} g(t) \, dt
$$

To facilitate the discussion, however, we rescale and translate the variable in the $i$-th segment by introducing the variable $x$ defined by

$$
x = \frac{2t - (t_{i+1} + t_i)}{t_{i+1} - t_i} \quad \text{or} \quad t = \frac{t_{i+1} + t_i}{2} + \frac{t_{i+1} - t_i}{2} x = t_i^{\text{mid}} + \frac{\Delta t_i}{2} x
$$

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

$$
\int_a^b g(t) \, dt = \sum_{i=1}^m \int_{t_i}^{t_{i+1}} g(t) \, dt = \sum_{i=1}^m \frac{\Delta t_i}{2} \int_{-1}^1 g\left(t_i^{\text{mid}} + \frac{\Delta t_i}{2} x\right) \, dx = \sum_{i=1}^m \frac{\Delta t_i}{2} \int_{-1}^1 f_i(x) \, dx
$$
where \( f_i(x) = g(t_i^{\text{mid}} + \Delta t_i \cdot x/2) \). In essence, then, we must evaluate an integral of the form

\[
\int_{-1}^{1} f(x) \, dx
\]

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

\[
\int_{-1}^{1} f(x) \, dx = \sum_{k=1}^{N} w_k f(x_k)
\]

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 2\( N \) 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

\[
\int_{-1}^{1} f(x) \, dx = w_1 f(x_1) + w_2 f(x_2)
\]

We choose \( w_1, w_2, x_1 \) and \( x_2 \) so that the formula gives the correct answer for the special cases

\[
\begin{align*}
\text{f}(x) = 1, \text{ yielding that } & \int_{-1}^{1} dx = 2 = w_1 + w_2 \quad (13.89) \\
\text{f}(x) = x, \text{ yielding that } & \int_{-1}^{1} x \, dx = 0 = w_1 x_1 + w_2 x_2 \quad (13.90) \\
\text{f}(x) = x^2, \text{ yielding that } & \int_{-1}^{1} x^2 \, dx = \frac{2}{3} = w_1 x_1^2 + w_2 x_2^2 \quad (13.91) \\
\text{f}(x) = x^3, \text{ yielding that } & \int_{-1}^{1} x^3 \, dx = 0 = w_1 x_1^3 + w_2 x_2^3 \quad (13.92)
\end{align*}
\]

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

\[
\begin{align*}
w_1 x_1 & = -w_2 x_2 \\
w_1 x_1^3 & = -w_2 x_2^3
\end{align*}
\]

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

\[
\frac{2}{3} = (w_1 + w_2)x_2^2 = 2x_2^2 \quad \implies \quad x_2^2 = \frac{1}{3}
\]

from which we also conclude that \( x_1^2 = 1/3 \). Thus,

\[
\begin{align*}
x_1 & = -\frac{1}{\sqrt{3}} \quad \text{and} \quad x_2 = \frac{1}{\sqrt{3}}
\end{align*}
\]

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.\(^{10}\)

Returning to the original function and variables as laid out in Eq. (13.85), we finally find that Eq. (13.96) supports the expression

\[
\int_a^b g(t) \, dt = \sum_{i=1}^{m} \frac{\Delta t_i}{2} \left[ g\left( t_{i+\text{mid}}^m - \frac{\Delta t_i}{2} \right) + g\left( t_{i+\text{mid}}^m + \frac{\Delta t_i}{2} \right) \right]
\]

\[
= \sum_{i=1}^{m} \frac{\Delta t_i}{2} \sum_{j=1}^{2} w_j g \left( t_{i+\text{mid}}^m + \frac{\Delta t_i}{2} \right)
\]

(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

\[
\int_{-1}^{1} f(x) \, dx = 0.23692689 f(-0.90617985) + 0.47862867 f(-0.53846931)
\]

\[
+ 0.56888889 f(0.00000000) \quad \text{(13.98)}
\]

\[
+ 0.47862867 f(0.53846931) + 0.23692689 f(0.90617985)
\]

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

\[
w_1 = 0.23692689 \quad x_1 = -0.90617985
\]

\[
w_2 = 0.47862867 \quad x_2 = -0.53846931
\]

\[
w_3 = 0.56888889 \quad x_3 = +0.00000000 \quad \text{(13.99)}
\]

\[
w_4 = 0.47862867 \quad x_4 = +0.53846931
\]

\[
w_5 = 0.23692689 \quad x_5 = +0.90617985
\]

Further, returning to the original variable, we note that

\[
\int_a^b g(t) \, dt = \sum_{i=1}^{m} \frac{\Delta t_i}{2} \sum_{j=1}^{5} w_j g \left( t_{i+\text{mid}}^m + \frac{\Delta t_i}{2} \right)
\]

(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 \((2/((1-x_i^2))(dL_2(x_i)/dx)^2)\). More generally, for an \(N\)-point Gaussian integration, we would discover that\(^{11}\)

\[
L_N(x_i) = 0 \quad \text{and} \quad w_i = \frac{2}{(1-x_i^2)(dL_N(x_i)/dx)^2}
\]

(13.101)

and that

\[
\left| \int_{-1}^{1} f(x) \, dx - \sum_{k=1}^{N} w_k f(x_k) \right| = \frac{2^{2N+1} (N!)^4}{(2N+1)((2N)!)^3} \left| \frac{d^{2N} f}{dx^{2N}} \right|_{x=\xi}
\]

(13.102)

\(^{10}\)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.

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*.\(^{12}\)

### 13.7 Evaluating Integrals Numerically with MATLAB

*Note:* All MATLAB program (.m) files referred to in this chapter are available in the directory `\$HEAD/matlab`, 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 MATLAB’s default search path. If so, the files can be found by MATLAB 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 MATLAB’s elementary commands or, more simply, by invoking one of the built-in routines `trapz`, `quad`, and `quadl`. We mention that MATLAB also provides a routine `quadl` for evaluating two-dimensional integrals but we invite the reader to study the MATLAB manuals for information on this additional capability.

#### 13.7.1 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

\[
I = \text{erf}(1) = \frac{2}{\sqrt{\pi}} \int_0^1 e^{-x^2} \, dx
\]

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\(^{13}\)

\[
\begin{align*}
>> & \text{format long} & \text{Set to display all digits.} \\
>> & n = 4; & \text{Set number of segments.} \\
>> & dx = 1.0/n; & \text{Set step size.} \\
>> & x = [0.0 : dx : 1.0]; & \text{Set values of } x. \\
>> & f = 2.0*exp(-x.^2)/sqrt(pi); & \text{Evaluate integrand.} \\
>> & y = 0.5*f(1); & \text{Initialize accumulator to first term.} \\
>> & \text{for } i = 2:n \text{ } y=y+f(i); \text{ end} & \text{Add middle terms.} \\
>> & y = y + 0.5*f(n+1); & \text{Add last term.} \\
>> & dx*y & \text{Rescale; display result.}
\end{align*}
\]

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 13.2. 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.

\(^{12}\)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 \(L_l(x)\) 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.

\(^{13}\)Alternatively, as long as \(n \geq 2\), the sixth, seventh, and eighth statements in this code can be replaced with the single statement \(y = 0.5*f(1) + \text{sum}( f(2:n) ) + 0.5*f(n+1)\).
Table 13.2: Values of \( \text{erf}(1.0) \) obtained by the trapezoidal rule and a user-constructed program in MATLAB. 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>0.771743332258054</td>
<td>64</td>
<td>0.842683902044948</td>
<td>4096</td>
<td>0.84270078882599</td>
</tr>
<tr>
<td>2</td>
<td>0.825262955596749</td>
<td>128</td>
<td>0.842696570249296</td>
<td>8192</td>
<td>0.84270079198783</td>
</tr>
<tr>
<td>4</td>
<td>0.838367777441205</td>
<td>256</td>
<td>0.84269973726221</td>
<td>16384</td>
<td>0.84270079219853</td>
</tr>
<tr>
<td>8</td>
<td>0.841619221244768</td>
<td>512</td>
<td>0.842700529031443</td>
<td>32768</td>
<td>0.842700792585281</td>
</tr>
<tr>
<td>16</td>
<td>0.842430505490233</td>
<td>1024</td>
<td>0.842700726970153</td>
<td>65536</td>
<td>0.84270079336114</td>
</tr>
<tr>
<td>32</td>
<td>0.842633227681257</td>
<td>2048</td>
<td>0.84270076454825</td>
<td>131072</td>
<td>0.842700792945679</td>
</tr>
</tbody>
</table>

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 or of shifting, for example, to quadruple precision arithmetic, 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

\[
S_2 = \frac{4 * T_2 - T_1}{3} = \frac{4 * 0.825262955596749 - 0.771743332258054}{3} = 0.843102830042981 \quad (13.104)
\]

and

\[
S_4 = \frac{4 * T_4 - T_2}{3} = \frac{4 * 0.838367777441205 - 0.825262955596749}{3} = 0.842736051389357 \quad (13.105)
\]

Continuing this process of Richardson extrapolation, we find the values \( S_8 = 0.842703035845956 \), \( S_{16} = 0.842700933572054 \), 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

\[
y = f * \text{transpose}(w)
\]

\(^{14}\)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 \( y = f * \text{transpose}(w) \).
Set number of segments.
Set step size.
Set values of \( x \).
Evaluate integrand.
Set first and last weights to 1.
Set proper weights to 4.
Set remaining weights to 2.
Initialize accumulator.
Add each strip with proper weight, suppressing output.
Rescale; display result.

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 MATLAB statements

```matlab
>> format long
>> pts = [-1.0/sqrt(3.0) 1.0/sqrt(3.0)];
>> w = [1.0 1.0];
>> a = 0.0; b = 1.0;
set evaluation points \( x_1 \) and \( x_2 \).
set weights \( w_1 \) and \( w_2 \).
set limits.
```

to set the evaluation points, weights, and limits. Then we would execute the statements

```matlab
>> m = 4;
>> dx = (b-a)/m;
>> x = [a : dx : b];
>> for i=1:m xmid(i)=(x(i)+x(i+1))/2.0; end
set number of segments.
set width of each segment.
set values dividing segments.
evaluate midpoints.
```

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

```matlab
>> 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 0.84244189522547, 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
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

```plaintext
>> m = 4;                       Set number of segments.
>> dx = (b-a)/m;                Set width of each segment.
>> x = [a : dx : b];            Set values dividing segments.
>> for i=1:m xmid(i)=(x(i)+x(i+1))/2.0; end;   Evaluate midpoints.
```

Finally, we evaluate the sum, multiply with an overall factor whose inclusion was postponed, and display the result with the statements

```plaintext
>> 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)
```

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 $0.842700789922891$—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!

### 13.7.2 The Functions trapz, quad, and quadl

MATLAB contains several built-in commands for the numerical evaluation of integrals. The first—`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

```plaintext
>> x = linspace( 0.0, 1.0, 17 );   Set values of independent variable.
>> y = 2.0*exp(-x.^2)/sqrt(pi);    Set values of dependent variable.
>> I = trapz( x, y )               Evaluate integral.
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 second method—`quad`—uses an adaptive Simpson’s rule, the third—`quadl`—uses recursive and adaptive Lobatto quadrature. Before invoking any of these routines, we must create a function M-file defining the integrand as a function of a single scalar argument (the integration variable). The construction of such files is described in Section 3.9. For the integral in Eq. (13.41), for example, the appropriate function M-file is listed in Table 13.3.

Having stored the M-file in the user’s current directory\(^{15}\) 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

\(^{15}\)See Section 3.18.2 for ways to point MATLAB to the proper directory.
### 13.7. EVALUATING INTEGRALS NUMERICALLY WITH MATLAB

Table 13.3: The MATLAB function `gausint.m`.

```matlab
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);
end
```

```matlab
q = quad( @function_name, a, b, tol )
```

in the newer—and now preferred—syntax or the general form

```matlab
q = quad( 'function_name', a, b, tol )
```

in the older syntax.\(^\text{16}\) In either form, `function_name`, which is mandatory, identifies the file in which the integrand is defined; \(a\) and \(b\)—also mandatory—are the lower and upper limits of integration; and `tol`, which is optional, is a single number specifying the desired absolute accuracy (default \(10^{-6}\)).\(^\text{17}\) The command `quad` returns the value of the integral in the specified variable, here \(q\).

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\(^\text{18}\)

```matlab
>> format long
>> q = quad( @gausint, 0.0, 1.0, 1.0e-3 )
q = 0.842700847014872
>> q = quad( @gausint, 0.0, 1.0 )
q = 0.842700847014872
>> q = quad( @gausint, 0.0, 1.0, 1.0e-9 )
q = 0.842700792955864
```

The first two of these values agree in all digits though, because the tolerance for the first value was set to \(10^{-3}\), we have no reason to believe that value beyond the third digit save that it agrees to the sixth digit with the second value—a value we would from its tolerance presume to be accurate to the sixth digit. On the basis of these first two values, we conclude that \(I = 0.842701\) to six digits.

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 `quadl` is identical to that described above for `quad`. Thus, the statements

\(^{16}\)We will use the newer form. If you are working with an older version of MATLAB, however, you may have to use the older form.

\(^{17}\)This default applies to versions after MATLAB5. In MATLAB5, `tol` was either a single number specifying the desired fractional accuracy (default \(10^{-3}\)) or a two-component vector whose first component specified the fractional accuracy and whose second component specified an absolute accuracy (default 0.0).

\(^{18}\)Values returned by MATLAB may well vary from version to version in digits less significant than those determined by the specified tolerance. Those presented here were returned by Version 7.14.0.739 (R2012a).
Table 13.4: The MATLAB function moment.m.

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;
end

>> q = quadl( @gausint, 0.0, 1.0 )
q = 0.842700794276711
>> q = quadl( @gausint, 0.0, 1.0, 1.0e-9 )
q = 0.842700792949715

use quadl to evaluate the integral to the default absolute tolerance of $10^{-6}$ and the stricter absolute tolerance $10^{-9}$. With quadl the two values differ beyond the eighth decimal place.

13.7.3 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 13.4 and store it in a file named moment.m. Then, the simple statements\(^{19}\)

\[
\begin{align*}
\text{>> format long} \\
\text{>> I = quad( @moment, 0.0, 1.0, 1.0e-3 )}  \\
\text{I = 0.247675864778771} \\
\text{>> I = quad( @moment, 0.0, 1.0 )}  \\
\text{I = 0.249995393009592} \\
\text{>> I = quadl( @moment, 0.0, 1.0, 1.0e-3 )}  \\
\text{I = 0.249914272120002} \\
\text{>> I = quadl( @moment, 0.0, 1.0 )}  \\
\text{I = 0.249999932876263}
\end{align*}
\]

evaluate the integral to one part in $10^3$ and one part in $10^6$ with quad and 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, each result (except the first) is 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 trapz with the statements

\(^{19}\)Again, values may differ from version to version. Those reported here were returned by Version 6.1.0.450 (R12.1).
13.7. EVALUATING INTEGRALS NUMERICALLY WITH MATLAB

```matlab
>> n = 10;
>> x = linspace( 0.0, 1.0, n+1 );
>> y = 4.0 * x.^2 .* sqrt(1.0-x.^2)/pi;
>> I = trapz( x, y )
I = 0.238496332979749
>> format short
```

to evaluate the desired integral via the trapezoidal rule with 10 segments in the interval 0 ≤ x ≤ 1.

We can repeat this process for larger values of n, finding the values

<table>
<thead>
<tr>
<th>n</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.238496332979749</td>
</tr>
<tr>
<td>100</td>
<td>0.249626708043829</td>
</tr>
<tr>
<td>1000</td>
<td>0.249988166015548</td>
</tr>
<tr>
<td>10000</td>
<td>0.249999925683605</td>
</tr>
<tr>
<td>100000</td>
<td>0.249999998162782</td>
</tr>
<tr>
<td>1000000</td>
<td>0.24999999999625674</td>
</tr>
</tbody>
</table>

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 0.25000.20

13.7.4 Quantum Probabilities

In Section 13.7.2, 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

\[ P(|x| > |x_{\text{turn}}|) = 1.0 - \text{erf}(1.0) = 1.0 - 0.842701 = 0.157299 \]  

(13.106)

We conclude that, in something over 15% of the measurements, the quantum oscillator will be found in the classically forbidden region!

13.7.5 Integrals as Functions of the Upper Limit

To evaluate an integral as a function of a parameter, `quad` or `quadl` 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

\[ g(x) = \int_0^x f(t) \, dt \]  

(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 ≤ x ≤ b`, we invoke the MATLAB statements21

---

20 Remember that MATLAB does its internal arithmetic in double precision, so we have probably not yet begun to be affected by problems from internal roundoff.

21 We might, of course, use `quadl` instead of `quad` in this example.
290 CHAPTER 13. EVALUATING INTEGRALS

\[ N = \langle \text{appropriate value} \rangle; \]
\[ x = \text{linspace}( a, b, N+1 ); \]
\[ \text{for } i = 1:N+1 \]
\[ \quad g(i) = \text{quad}( @\text{integ}, 0, x(i) ); \]
\[ \text{end;} \]

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 \( \text{plot}(x,g) \) will generate a graph of \( g \) versus \( x \).\(^{22}\)

13.7.6 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 \( \text{gausint.m} \) defined in Section 13.7.2 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 MATLAB statements

\[ >> N = 30; \]
\[ >> x = \text{linspace}( 0.0, 3.0, N+1 ); \]
\[ >> \text{for } i = 1:N+1 \]
\[ \quad \text{erf}(i) = \text{quad}( @\text{gausint}, 0, x(i) ); \]
\[ \text{end} \]
\[ >> \text{plot}( x, \text{erf}, '\text{LineWidth}', 4, \ldots \]
\[ \quad \text{'Color', 'black'}) \]
\[ >> \text{title('Error Function', 'FontSize', 20 )} \]
\[ >> \text{xlabel('}x', \text{'FontSize', 16 )} \]
\[ >> \text{ylabel('}\text{erf}(x)', \text{'FontSize', 16 )} \]
\[ >> \text{grid on} \]

The resulting graph is shown in Fig. 13.10.

13.7.7 The Cornu Spiral

To determine the Cornu spiral, we use \( \text{quad} \) 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 listed in Table integ:spiralMAT to provide the two integrands, storing them in the user’s current directory with the names \( \text{spiralc.m} \) and \( \text{spirals.m} \). 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

\[ >> N = 100; \]
\[ >> u = \text{linspace}( 0, 5.0, N+1 ); \]
\[ >> \text{for } i = 1:N+1 \]
\[ \quad \text{CP}(i) = \text{quad}( @\text{spiralc}, 0, u(i) ); \]
\[ \quad \text{SP}(i) = \text{quad}( @\text{spirals}, 0, u(i) ); \]
\[ \quad \text{CN}(N+2-i) = -\text{CP}(i); \]
\[ \quad \text{SN}(N+2-i) = -\text{SP}(i); \]
\[ \text{end} \]

Finally, we assemble the variables to be plotted and plot the graph with the statements

\[ \text{This sequence of statements actually is computationally inefficient. We might increase the efficiency by recognizing, for example, that } g(x + \Delta x) = g(x) + \int_x^{x+\Delta x} f(t) \, dt \text{ and obtain integrals for larger } x \text{ by adding an appropriate increment to already evaluated integrals for smaller } x. \text{ For our present purposes, that approach unnecessarily complicates the algorithm of evaluation.} \]
13.7. Evaluating Integrals Numerically with MATLAB

Figure 13.10: The error function.

Table 13.5: The MATLAB functions spiralc.m and spirals.m.

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 );
end

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 );
end

% Concatenate values for positive and negative limits.
% Plot graph of S versus C.
% Adjust axes.
% Turn on grid.
% Label graph.

>> C = [CN, CP];
>> S = [SN, SP];
>> plot( C, S, 'LineWidth', 2, ... 'Color', 'black' )
>> axis( [-1.0 1.0 -1.0 1.0 ] )
>> axis square
>> grid on
>> title( 'Cornu Spiral', 'FontSize', 20 )
>> xlabel( 'C(u)', 'FontSize', 16 )
>> ylabel( 'S(u)', 'FontSize', 16 )
Figure 13.11: The Cornu spiral.

The resulting graph is shown in Fig. 13.11.

13.7.8 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 quad or quadl 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 3.9. This objective can more conveniently be accomplished by exploiting additional optional arguments to quad or quadl. The full structure of statements involving these commands is

\[
q = \text{quad}( \text{@function\_name}, a, b, \text{tol}, \text{trace}, p1, p2, \ldots )
\]

\[
q = \text{quadl}( \text{@function\_name}, a, b, \text{tol}, \text{trace}, p1, p2, \ldots )
\]

in the newer syntax and

\[
q = \text{quad}( \text{function\_name}, a, b, \text{tol}, \text{trace}, p1, p2, \ldots )
\]

\[
q = \text{quadl}( \text{function\_name}, a, b, \text{tol}, \text{trace}, p1, p2, \ldots )
\]

in the older syntax. Here—beyond the quantities already defined—a non-zero value for \text{trace} will produce a graphical display showing the points at which the function is evaluated as the integration proceeds, and \text{p1}, \text{p2}, \ldots are parameters that will be passed by quad or quadl to the function defining the integrand, provided that function is defined with those parameters as additional arguments. To accept the default values of \text{tol} and/or \text{trace} while specifying parameters, simply set either/both to the empty matrix [].

\footnote{This sequence of statements actually has two glitches. First, it is computationally inefficient in the way described in footnote 22. Second, and less significantly, the values of \(S(0)\) and \(C(0)\) appear twice in the vectors finally plotted.}
13.7. EVALUATING INTEGRALS NUMERICALLY WITH MATLAB

Table 13.6: The MATLAB function rings.m.

```matlab
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 );
end
```

To invoke this feature, we must then 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

```matlab
function y = function_name( x, p1, p2, p3, ... )
% Line to be scanned by lookfor.
% Explanatory comments.
Statements to evaluate integrand, using p1, p2, p3, ... for the parameters.
y = integrand;
```

The quantities `function_name`, `x`, and `integrand` have the same meaning as before, and the file
must be stored with the name `function_name.m`.

Second, having defined the function M-file in this new way and stored it in the current directory,
we invoke MATLAB to evaluate and display the desired integral by using `quad` with the statements

```matlab
p1 = ??;, p2 = ??;, p3 = ??;
q = quad( @function_name, a, b, [], [], ...
          p1, p2, p3 )
```

Note that the `names` of the parameters need *not* be the same at command level and 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.

13.7.9 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 in Table 13.6 and store it
with the name `rings.m`, then invoke MATLAB to evaluate and display the desired integral by using `quad` with the statements

```matlab
q = quad( @rings, a, b, [], [],...
          s )
```

Since we seek to explore this integral as a function of `s` over, say, 0.0 ≤ `s` ≤ 4.0, we must generate
a vector of values of `s` and then, in a loop, invoke `quad` once for each element in that vector. We
might use the statements

```
24 We could, of course, have used `quadl` instead of `quad` in this example.
```
Figure 13.12: The potential in the midplane of a charged ring.

\[ \text{potential in midplane} \]

\[ V(s) \text{ (in units of } V(0)) \]

\[ s \text{ (in units of } a) \]

```matlab
>> s = [ 0.0 : 0.1 : 4.0 ];
>> N = length(s);
>> for i = 1:N
  V(i) = quad(@rings, 0.0, 2.0*pi, [], [], s(i));
end
```

Specify values of parameter. Set \( N \) to number of values in \( s \). Evaluate and store integral for each value of the parameter.

```matlab
>> plot( s, V, 'LineWidth', 4, 'Color', 'black' )
```

Plot graph. 

```matlab
>> axis([0.0 4.0 0.0 1.0])
>> title('Potential in Midplane', 'FontSize', 20)
>> xlabel('s (in units of } a)', 'FontSize', 16)
>> ylabel('V(s) (in units of } V(0))', 'FontSize', 16)
>> grid on
```

Adjust axis limits. Title graph. Label axes. Turn grid on.

The resulting graph is shown in Fig. 13.12.

### 13.10 Evaluating Integrals Numerically with MAXIMA

In addition to its routines for addressing integrals symbolically, MAXIMA contains several commands for numerical evaluation (sometimes called *quadrature*) of one-dimensional definite integrals. Among the available commands are the following:

- \texttt{quad_qag( } f(x), x, a, b, key, epsrel=..., epsabs=..., limit=... \texttt{ )}, which integrates a general function over a finite interval by globally adaptive Gaussian integration.

- \texttt{quad_qags( } f(x), x, a, b, epsrel=..., epsabs=..., limit=... \texttt{ )}, which adds extrapolation to the procedure invoked by \texttt{quad_qag}. 
• **quad_qagi**(*f(x), x, a, b, epsrel=..., epsabs=..., limit=...*), which integrates a general function over an infinite or semi-infinite interval by mapping the interval into a finite interval and then invoking **quad_qags**.

The arguments in these functions are

- \( f(x) \): the function to be integrated, which can be either the full function, e.g., \( \sqrt{1+x^2} \) or the label to which the function has been bound.
- \( x \): the integration variable.
- \( a \): the lower limit.
- \( b \): the upper limit.
- \( key \): (only in **quad_qag**) an integer between 1 and 6, inclusive, specifying the order of the Gaussian integration rule to be invoked.
- \( \text{epsrel} \): desired relative error; default is \( 10^{-8} \) in double precision.
- \( \text{epsabs} \): desired absolute error; default is 0.
- \( \text{limit} \): maximum number of subintervals into which the adaptive procedure will divide the interval \( a \leq x \leq b \); default is 200.

At least one of \( a \) and \( b \) must be infinite in **quad_qagi**. Note also that the three keyword arguments \( \text{epsrel}, \text{epsabs}, \text{and} \text{limit} \) are optional, can be presented in any order, and take the form of **keyword** = **value**. Descriptions of additional routines for evaluating Fourier transforms, weighted integrals, and Cauchy principal values of integrals can be found in the MAXIMA manuals.

These commands return a four-component list consisting of, in order, an approximation to the integral, the estimated absolute error of the approximation, the number of times the integrand was evaluated, and an error code. The error code will have one of the values

- 0: if no problems were detected in the evaluation.
- 1: if the maximum number of subintervals was exceeded.
- 2: if excessive roundoff error was detected.
- 3: if very bad behavior occurs.
- 6: if the input function is invalid.

With these features of MAXIMA, we can evaluate our standard example in many ways. We might, for example use the statement

```
(%i1) float( integrate( 2*exp(-t^2)/sqrt(%pi), t, 0, 1 ) );
(%o1)  0.8427007929497148
```

which will evaluate the integral analytically [obtaining \( \text{erf}(1) \)], evaluate the error function of argument 1, and finally display the result, presumably to double-precision accuracy. Alternatively, we might accept defaults for all parameters and use the statement

```
(%i2) quad_qag( 2*exp(-t^2)/sqrt(%pi), t, 0, 1, 2);
(%o2)  [0.8427007929497151, 9.35585 \ldots \times 10^{-15}, 21, 0]
```

which yields the value 0.842700... to a claimed absolute error of \( 9 \times 10^{-15} \) with 21 evaluations of the integrand and no problems detected in the evaluation. If we increase the order of the integration but still accept the defaults for the parameters, we find that

```
(%i3) quad_qag( 2*exp(-t^2)/sqrt(%pi), t, 0, 1, 4);
(%o3)  [0.8427007929497151, 9.35585 \ldots \times 10^{-15}, 41, 0]
```

These results are all in agreement to the 15th decimal place.
13.10.1 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

\[ P(|x| > |x_{\text{turn}}|) = 1.0 - \text{erf}(1.0) = 1.0 - 0.8427008 = 0.1572992 \]  \hspace{1cm} (13.108)

which has been rounded to seven digits after the decimal point. We conclude that in a bit over 15\% of the measurements, the quantum oscillator will be found in the classically forbidden region!

13.10.2 The Error Function

To define and integrate 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}
(%i1) gaussian : 2.0 * exp(-t^2) / sqrt(%pi)$
(%i2) for j : 0 thru 30 do (q[j]:quad_qag(gaussian, t, 0, j/10, 2), i[j]:j/10)$
(%i3) [ i[0], q[0] ];
(%o3) [ 0,0.0,0.0,21,0 ]

;  \hspace{1cm} \text{Check evaluation for upper limit = 0}

(%i4) [ i[10], q[10] ];
(%o4) [ 1,0.8427...,9.355... \times 10^{-15},21,0 ]

;  \hspace{1cm} \text{Check evaluation for upper limit = 1}
\end{verbatim}

The arrays \( q \) and \( i \) created by the statement on line \%12 must then be converted into lists with the statements

\begin{verbatim}
(%i5) i1 : makelist( i[v], v, 0, 30 )$
(%i6) q1 : makelist( q[v][1], v, 0, 30 )$
\end{verbatim}

before the graph of Fig. 13.13 can be created. Here, the notation \( q[v][1] \) in \%16 extracts the \( v \)-th four-component list from \( q \) and then the first component of that four-component list to generate in \( q1 \) a list whose \( i \)-th component is solely the evaluation of the integral with an upper limit equal to the \( i \)-th component of \( i1 \). The resulting integral as a function of its upper limit can be graphed with the statement

\begin{verbatim}
(%i9) plot2d( [discrete, i1, q1], [xlabel, "x"], [ylabel, "erf(x)"]$, [title, "Error Function"], [style, [lines, 4]], [color, black], [grid2d, true] )$
\end{verbatim}

Though it will make little difference in the present example, integration will proceed more quickly if the function to be integrated is compiled before the integrals are evaluated and the above statements generating the graph are executed. To that end, we would define the function and compile it with the statements

\begin{verbatim}
(%i1) gaussian(t) := 2.0 * exp(-t^2) / sqrt(%pi)$
(%i2) compile(gaussian)$
(%i2) for j : 0 thru 30 do (q[j]:quad_qag(gaussian, t, 0, j/10, 2), i[j]:j/10)$
\end{verbatim}
13.15. EXERCISES

13.15.3 The Off-Axis Electrostatic Potential of Two Rings

Using MAXIMA, 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 statements

\[
\begin{align*}
\text{(i1)} & \quad \text{mode}\_\text{declare( [ function(rings), phi, s ], float )} \\
\text{(i2)} & \quad \text{rings(phi) := sqrt(2.0) / (2.0*%pi) / sqrt(2.0-2.0*s*cos(phi)+s^2)} \\
\text{(i3)} & \quad \text{compile(rings)} \\
\text{(i4)} & \quad \text{for ss: 0 thru 30 do} \\
& \quad \quad \text{s1 : makelist( i[k], k, 0, 30 )} \\
& \quad \quad \text{v1 : makelist( v[k][1], k, 0, 30 )} \\
\text{(i5)} & \quad \text{plot2d( [discrete, s1, v1], [xlabel, "s"], [ylabel, "V(s)"],} \\
& \quad \quad \text{[title, "Midplane Potential"], [style, [lines, 4]],} \\
& \quad \quad \text{[color, black], [grid2d, true], [y, 0.0, 1.0] })
\end{align*}
\]

make sure MAXIMA understands that certain variables stand for floating point values, define the integrand, integrate the function as the parameter \( s \) varies from 0.0 to 3.0 in steps of 0.1, and produce and label 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 $\mathbf{x}$ and variance $\sigma^2$, defined by

$$\mathbf{x} = \lim_{b \to \infty} \int_{-b}^{b} x p(x) \, dx \quad \text{and} \quad \sigma^2 = \lim_{b \to \infty} \int_{-b}^{b} (x - \mathbf{x})^2 p(x) \, dx$$

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 $\lambda = x/a$.

13.6. 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,
13.15. EXERCISES

what is the value of this time for the earth-sun system? Hint: Since the gravitational potential is $-GmM/x$, conservation of energy yields

$$\frac{1}{2}m\left(\frac{dx}{dt}\right)^2 - G\frac{mM}{x} = -G\frac{mM}{x_0} \implies \frac{dx}{dt} = -\sqrt{2GM} \sqrt{\frac{1}{x} - \frac{1}{x_0}}$$

(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

$$T_{\text{midhalf}} = \frac{1}{\sqrt{2GM}} \int_{x_0/4}^{3x_0/4} \left(\frac{1}{x} - \frac{1}{x_0}\right)^{-1/2} dx$$

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

$$P(r) = \frac{4}{a^3} \int_{r_0}^{r} e^{-2\rho/a} r'^2 dr' = 4 \int_{0}^{r/a} e^{-2\rho} \rho^2 d\rho$$

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

$$V(z) = \frac{\sigma}{2\epsilon_0} \left[\sqrt{a^2 + (z-ca)^2} - |z-ca| - \sqrt{a^2 + (z+ca)^2} + |z+ca|\right]$$

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, $r = z\hat{k}$, for which $z > a$ by setting up and evaluating the integral

$$V(0,0,z) = \frac{1}{4\pi\epsilon_0} \int \frac{\rho(r') \, dr'}{|r - r'|}$$

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 $z'$.

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

$$I_{ij} = \int \left[(x_1^2 + x_2^2 + x_3^2)\delta_{ij} - x_i x_j\right] dm$$
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

$$x_{mn} = \int_{-\infty}^{\infty} \psi_m^*(x) x \psi_n(x) \, dx \quad \text{and} \quad p_{mn} = \int_{-\infty}^{\infty} \psi_m^*(x) \frac{\hbar}{i} \frac{\partial}{\partial x} \psi_n(x) \, dx$$

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

$$x_{mn} = \frac{16a}{\pi^2} (-1)^{(m+n+1)/2} \frac{mn}{(m^2 - n^2)^2}$$

$$p_{mn} = \frac{2i\hbar}{a} (-1)^{(m+n+1)/2} \frac{mn}{(m^2 - n^2)^2}$$

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

$$f_{\text{trunc}}(x) = \frac{a_0}{2} + \sum_{n=1}^{N} \left( a_n \cos \frac{n\pi x}{l} + b_n \sin \frac{n\pi x}{l} \right)$$

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

$$c_n = \frac{2n+1}{2} \int_{-1}^{1} f(x) L_n(x) \, dx$$

(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 $N = 0, 1, 2, 3, 4, 5, 6$. 

13.15. EXERCISES

13.15.2 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 \( v \). *Hint:* Re-express the integral using \( \sqrt{2kT/m} \) as the unit of velocity.

**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

\[
\begin{align*}
L_0(x) &= 1, & L_5(x) &= \frac{1}{8}(63x^5 - 70x^3 + 15x) \\
L_1(x) &= x, & L_6(x) &= \frac{1}{16}(231x^6 - 315x^4 + 105x^2 - 5) \\
L_2(x) &= \frac{1}{2}(3x^2 - 1), & L_7(x) &= \frac{1}{16}(429x^7 - 693x^5 + 315x^3 - 35x) \\
L_3(x) &= \frac{1}{2}(5x^3 - 3x), & L_8(x) &= \frac{1}{128}(6435x^8 - 12012x^6 + 6930x^4 - 1260x^2 + 35) \\
L_4(x) &= \frac{1}{2}(35x^4 - 30x^2 + 3), & L_9(x) &= \frac{1}{128}(12155x^9 - 25740x^7 + 18018x^5 - 4620x^3 + 315x)
\end{align*}
\]

13.16. Repeat the evaluations of the sample integrals in Section 13.7.2 with *quad*, *quad8*, and *quad1*, specifying a nonzero trace, which will reveal the number of evaluations of the function and the sequence in which those evaluations are performed. Write a paragraph describing your observations.

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

\[
\begin{align*}
\psi_1(x) &= \sqrt{2} \left( \frac{m\omega}{\pi\hbar} \right)^{1/4} ye^{-y^2/2}; \\
\psi_2(x) &= \sqrt{2} \left( \frac{m\omega}{\pi\hbar} \right)^{1/4} (2y^2 - 1) e^{-y^2/2}
\end{align*}
\]

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. 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.
13.20. Planck’s black body radiation law gives the expression
\[ I(\nu_2, \nu_1) = \frac{8\pi h}{c^3} \int_{\nu_1}^{\nu_2} \frac{\nu^3}{e^{h\nu/kT} - 1} d\nu \]
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. \( \text{Hint: One way to approach this exercise would be to choose a reference frequency } \nu_0 \text{ arbitrarily (say } 10^{14} \text{ Hz) and recast the integral on the dimensionless variable } s = \nu/\nu_0. \) Examination of \( I \) in units of \( 8\pi h\nu_0^4/c^3 \) as a function of \( T \) in units of \( h\nu_0/k \) 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
\[ p(t) = \frac{1}{\sqrt{2\pi} \sigma} e^{-t^2/2\sigma^2} \]
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 \( P(-x, x) = \text{erf}(x/\sqrt{2}\sigma) \), and then evaluate \( P(-\sigma, 2\sigma) \), \( P(-3\sigma, 3\sigma) \) 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? \( \text{Hint: Since the gravitational potential is } -GmM/x, \text{ conservation of energy yields} \)
\[ \frac{1}{2}m \left( \frac{dx}{dt} \right)^2 - G\frac{mM}{x} = -G\frac{mM}{x_0} \quad \implies \quad \frac{dx}{dt} = -\sqrt{2GM} \sqrt{\frac{1}{x} - \frac{1}{x_0}} \]
(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
\[ T_{midhalf} = \frac{1}{\sqrt{2GM}} \int_{x_0/4}^{3x_0/4} \left( \frac{1}{x} - \frac{1}{x_0} \right)^{-1/2} dx \]
\( \text{Hint: The evaluation will be simpler if you begin by recasting the problem in dimensionless terms, expressing lengths in units of } x_0 \text{ and times in units of } \sqrt{x_0^4/(2GM)}. \) 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. \( \text{Optional: See if you can develop a means to determine the time required for the first half of the journey, which unfortunately—for numerical approaches—involves a convergent but improper integral.} \)

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 not exploit a priori knowledge of the exact value. \( \text{Hint: Before evaluating the integral, introduce the dimensionless variable } s = x/a \text{ and note that the result actually doesn’t depend on } a, \text{ so there is but one number to determine.} \)
13.15. EXERCISES 303

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

$$P(r) = \frac{4}{a^3} \int_0^r e^{-2r'/a} r'/dr' = 4 \int_0^r e^{-2\rho^2} \rho^2 d\rho$$

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

$$K(k) = \int_0^{\pi/2} \frac{d\phi}{(1 - k^2 \sin^2 \phi)^{1/2}} ; \quad E(k) = \int_0^{\pi/2} (1 - k^2 \sin^2 \phi)^{1/2} d\phi$$

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

$$\omega t = \int_0^\beta \frac{d\phi}{(1 - k^2 \sin^2 \phi)^{1/2}}$$

where, with $l$ the length of the pendulum and $g$ the acceleration of gravity, $\omega = \sqrt{g/l}$, $k = \sin(\alpha/2)$, and $\beta = \sin^{-1}(\sin(\alpha/2))/l)$. 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 $\omega t$ 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

$$\frac{1}{x} J_1(x) = \frac{2}{\pi} \int_0^1 (1 - u^2)^{1/2} \cos(xu) \, du$$

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

$$V(r, \phi, z) = \frac{Q}{4\pi \epsilon_0 a} = \frac{1}{\pi} \int_0^{\pi} \left( 1 - 2 \frac{r^2}{a^2} \cos \phi' + \frac{r^2}{a^2} + \frac{z^2}{a^2} \right)^{-1/2} \, d\phi'$$

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

$$\frac{B(x, z)}{\mu_0 I/2\pi a} = a^2 \int_0^{\pi} \frac{z \cos \phi'}{|x^2 + z^2 + a^2 - 2ax \cos \phi'|^{3/2}} \, d\phi'$$

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).
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

$$f(x) = 0 \quad (14.1)$$

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.\footnote{The graphs could, of course, be produced in any number of ways. Except for Fig. 14.2 (which was produced with 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.}

\section*{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.
**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

$$V(x) = E \quad \text{or} \quad V(x) - E = 0$$

(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

$$V(x) = \frac{x^3}{10000} + \frac{x^2}{200} - \frac{x}{500} - \frac{1}{2}$$

(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

$$-50.00 < x_1 < -47.50 \quad ; \quad -12.50 < x_2 < -10.00 \quad ; \quad 7.50 < x_3 < 10.00$$

(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

$$F = 0 \quad \implies \quad \frac{dV}{dx} = 0$$

(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

$$\frac{dV}{dx} = \frac{3x^2}{10000} + \frac{x}{100} - \frac{1}{500} = 0$$

(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

$$-35.0 < x_4 < -30.0 \quad \text{and} \quad -2.5 < x_5 < 2.5$$

(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 $V(x) = kx^2/2 + ax^3$ 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.
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\(^3\)

\[
\begin{align*}
    m \frac{d^2 x_1}{dt^2} &= -k x_1 + k'(x_2 - x_1) \\
    m \frac{d^2 x_2}{dt^2} &= -k x_2 - k'(x_2 - x_1)
\end{align*}
\]

To cast these equations in dimensionless form, we choose a unit of length $a$, set $x_i/a = \bar{x}_i$, introduce $\omega_0 = \sqrt{k/m}$, set $\bar{k} = k'/k$, and introduce $\bar{t} = \omega_0 t$ to find that

\[
\begin{align*}
    \frac{d^2 \bar{x}_1}{d\bar{t}^2} &= -(1 + \bar{k}) \bar{x}_1 + \bar{k} \bar{x}_2 \\
    \frac{d^2 \bar{x}_2}{d\bar{t}^2} &= \bar{k} \bar{x}_1 - (1 + \bar{k}) \bar{x}_2
\end{align*}
\]

Next, seeking sinusoidal (or simple harmonic) solutions, we suppose that

\[
\begin{align*}
    \bar{x}_1(\bar{t}) &= \bar{x}_{10} \cos \omega \bar{t} \\
    \bar{x}_2(\bar{t}) &= \bar{x}_{20} \cos \omega \bar{t}
\end{align*}
\]

\(^3\)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

\[
\begin{pmatrix} 1 + \kappa & -\omega \omega_0^2 \\ -\omega & 1 + \kappa - \omega \omega_0^2 \end{pmatrix} \begin{pmatrix} \pi_{10} \\ \pi_{20} \end{pmatrix} = 0
\] (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

\[
D(\omega) = \begin{vmatrix} 1 + \kappa & -\omega \omega_0^2 \\ -\omega & 1 + \kappa - \omega \omega_0^2 \end{vmatrix}
\]

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 \( \kappa = k'/k \), 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 \( \kappa \) 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 \( \kappa \), while the upper of the two roots increases steadily as \( \kappa \) 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

\[
x(t) = \frac{m v_0 \cos \theta}{b} \left( 1 - e^{-bt/m} \right)
\] (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

\[
\frac{R(\theta, \alpha)}{v_0/g} = \frac{\cos \theta (1 - e^{-\tau_1})}{\alpha} = \frac{\cos \theta}{\alpha(1 + \alpha \sin \theta)} \tau_1
\] (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

\[
0 = \tau - (1 + \alpha \sin\theta) \left[ 1 - \left( 1 - \tau + \frac{\tau^2}{2} \right) \right] = \tau + (1 + \alpha \sin\theta) \left[ -\tau + \frac{\tau^2}{2} \right] \\
= -\alpha \tau \sin\theta + \frac{\tau^2}{2} = 0 \quad \Rightarrow \quad \tau = 2\alpha \sin\theta
\]  

(14.20)

Then, substituting this result into Eq. (14.19), we find that

\[
\frac{R(\theta, \alpha \to 0)}{v_0^2/g} = \frac{\cos\theta}{\alpha(1 + \alpha \sin\theta)} (2\alpha \sin\theta) = \frac{2 \sin\theta \cos\theta}{1 + \alpha \sin\theta} \to 2 \sin\theta \cos\theta
\]  

(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

\[
V(x) = \begin{cases} 
\infty & x < 0 \\ 
-V_0 & 0 \leq x \leq a \\ 
0 & a \leq x 
\end{cases}
\]  

(14.22)

leads to the conclusion that the allowed energies \( E \) should satisfy the equation\(^5\)

\[
s \cot s = -\sqrt{c^2 - s^2}
\]  

(14.23)

where \( c^2 = 2ma^2V_0/\hbar^2 \) 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

\[
s^2 \cot^2 s = s^2 (\sin^{-2} s - 1) = c^2 - s^2 \quad \Rightarrow \quad \sin s = \pm \frac{s}{c}
\]  

(14.24)

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 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

\[
\sin s = \pm 0.04s
\]  

(14.25)

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 its roots lie where 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 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 upper sign in Eq. (14.25) the eight roots are bounded—crudely—by

\[
\begin{align*}
\text{root 1u} & : -0.5 < s < 0.5  \\
\text{root 2u} & : 2.5 < s < 3.5  \\
\text{root 3u} & : 6.0 < s < 7.0  \\
\text{root 4u} & : 8.5 < s < 9.5
\end{align*}
\]  

(14.26)

and that, for the lower sign, the nine roots are bounded—again crudely—by

\[
\begin{align*}
\text{root 1l} & : -0.5 < s < 0.5  \\
\text{root 2l} & : 3.0 < s < 4.0  \\
\text{root 3l} & : 5.5 < s < 6.5  \\
\text{root 4l} & : 9.5 < s < 10.5  \\
\text{root 5l} & : 11.5 < s < 12.5
\end{align*}
\]  

(14.27)

\(^5\)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, 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 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

$$f(x) = ax + b = 0 \quad \Rightarrow \quad x = -\frac{b}{a}$$

(14.28)

and that the quadratic formula

$$f(x) = ax^2 + bx + c = 0 \quad \Rightarrow \quad x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

$$= \frac{2c}{-b \mp \sqrt{b^2 - 4ac}}$$

(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

$$f(x) = \sin(kx) = 0 \quad \Rightarrow \quad x = \frac{n\pi}{k}, \quad n = 0, \pm 1, \pm 2, \ldots$$

(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

\[ \frac{a + x}{b + x^2} - c = 0 \quad \implies \quad a + x - c(b + x^2) = -cx^2 + x + a - cb = 0 \]  

converts the problem into one involving a quadratic polynomial. Beyond these few cases, very few roots can be found symbolically.

### 14.3 Finding Roots Symbolically with MAXIMA

The MAXIMA command for solving a wide variety of algebraic and transcendental equations is \texttt{solve}. It takes two arguments, the first giving the equation to be solved (or the function whose roots are desired) and the second specifying the variable for which a solution is desired. Thus, the statement

\[
\begin{align*}
\text{(i1)} \quad & \text{solve( a*x + b = 0, x )}; \\
\text{(o1)} & \quad \bigg[ x = -\frac{b}{a} \bigg]
\end{align*}
\]

returns the solution \( x = -b/a \) and the statement

\[
\begin{align*}
\text{(i2)} \quad & \text{solve( a*x^2 + b*x + c, x )}; \\
\text{(o2)} & \quad \bigg[ x = -\frac{\sqrt{b^2 - 4ac} + b}{2a}, x = \frac{\sqrt{b^2 - 4ac} - b}{2a} \bigg]
\end{align*}
\]

in which we have illustrated that MAXIMA assumes the omitted ‘\( = 0 \)’ at the end of the first argument, returns two solutions

\[ x = \frac{-b + \sqrt{b^2 - 4ac}}{2a} ; \quad x = \frac{-b - \sqrt{b^2 - 4ac}}{2a} \]  

(14.33)

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 that reduces all of the factors to polynomials of order no higher than quartic. Note also that \texttt{solve} returns the solution(s) as a list, even if there is only one solution.

### 14.3.1 Classical Turning Points

The turning points for the potential energy in Eq. (14.3) are readily found symbolically with the statements
\( V : x^3/10000 + x^2/200 - x/500 - 1/2 \)
\( \text{solve} \ (V=0, \ x) \)
\[
\begin{align*}
\text{soln} = & \ [x = \text{mess}, \ x = \text{mess}, \ x = \text{mess}] \\
\end{align*}
\]
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, generate a more interpretable solution by extracting and simplifying the real and imaginary parts with the statements\(^6\)

\( \text{map} \ (\text{rhs}, \ \text{realpart} (\text{float} (\text{soln}))) \)

\( \text{map} \ (\text{rhs}, \ \text{imagpart} (\text{float} (\text{soln}))) \)

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 the precision of the command \text{float} is on the order of \( 10^{-15} \), so the imaginary parts differ from zero almost certainly because of internal computer roundoff.

### 14.3.2 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

\( \text{diff} \ (V, \ x) \)
\( \text{solve} \ (\text{diff} (V, \ x) = 0, \ x) \)

\[
\begin{align*}
\text{soln} = & \ [x = -16\sqrt{10} + 50/3, \ x = 16\sqrt{10} - 50/3] \\
\end{align*}
\]

\( \text{float} \)

\[
\begin{align*}
[x = -33.53214752089802, \ x = 0.198814187564908] \\
\end{align*}
\]

finding first \( dV/dx \) and then finding the values of \( x \) at which \( dV/dx = 0 \). The values 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

\( \text{diff} \ (dV/dx, \ x) \)
\( \text{part} \ (\text{soln}, \ 1) \)

\[
\begin{align*}
\text{Vdx2} = & \ [-0.01011928851253881] \\
\text{Vdx2, part} (\ %o8, \ 2) = & \ [0.01011928851253881] \\
\end{align*}
\]

concluding that the first root \( x = -33.53214\ldots \) locates an unstable equilibrium \( (d^2V/dx^2 < 0; \text{local maximum in the potential energy}) \) while the second root \( x = 0.198814\ldots \) locates a stable equilibrium \( (d^2V/dx^2 > 0; \text{local minimum in the potential energy}) \).

\(^6\)Different versions of MAXIMA run on different platforms may produce slightly different results in the last digit or two.
14.3. 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 $\omega^2$, which we call $\omega_2$, and then take the square root of that solution with the statements:

\[
\begin{align*}
\text{(i1)} & \quad d := \omega_2^2 - 2*\omega_2*(1+\kappa) + 1 + 2*\kappa \\
\text{(i2)} & \quad \text{solve}( d = 0, \omega_2 ) \\
\text{(o2)} & \quad \left[ \omega_2 = 2\kappa + 1, \omega_2 = 1 \right] \\
\text{(i3)} & \quad \text{map}( \text{rhs}, \text{o2} ) \\
\text{(o3)} & \quad [2\kappa + 1, 1] \\
\text{(i4)} & \quad \omega := \sqrt{\%} \\
\text{(o4)} & \quad [\sqrt{2\kappa + 1}, 1]
\end{align*}
\]

Note—as must on physical grounds be the case—that the solutions obtained for $\omega^2$ at (o2) 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 increases steadily as $\kappa$—the strength of the coupling—increases, but that the second of the two frequencies is the same for all $\kappa$.

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 statements:

\[
\begin{align*}
\text{(i5)} & \quad \text{plot2d}( \omega, \left[ \kappa, 0, 4 \right], \left[ \text{xlabel}, \text{"kappa"} \right], \left[ \text{ylabel}, \text{"omega"} \right], \\
& \quad \quad \quad \quad \quad \left[ y, 0, 3 \right], \left[ \text{legend}, \text{false}, \right], \left[ \text{color}, \text{black}, \right], \left[ \text{style}, \left[ \text{lines}, 4 \right] \right] );
\end{align*}
\]

The resulting graph is shown in Fig. 14.6.

14.3.4 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 MAXIMA. We begin by defining the trajectory as given by Eqs. (14.15) and (14.16) with the statements:

\[
\begin{align*}
\text{(i1)} & \quad x := \frac{(m*v_0*\cos(\theta))/b}{1 - \exp(-b*t/m)} \\
\text{(i2)} & \quad y := \frac{-m*g*t/b+(m/b)*(v_0*\sin(\theta)+m*g/b)}{1 - \exp(-b*t/m)}
\end{align*}
\]

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 MAXIMA statements:

\[\text{At this point, for simplicity, we drop the overbar on } \pi.\]

\[\text{To facilitate visualization, some of the output in the following dialog with MAXIMA has been recast from the actual form presented by MAXIMA.}\]
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 $\omega_0 = \sqrt{k/m}$.

(%i3) xb : factor(taylor(x, b, 0, 1));
(%o3) $-\frac{t(bt - 2m) \cos \theta v_0}{2m}$

(%i4) yb : expand(taylor(y, b, 0, 1));
(%o4) $-\frac{bt^2 \sin \theta v_0}{2m} + t \sin \theta v_0 + \frac{bg t^3}{6m} - \frac{gt^2}{2}$

If we notice some of the features of these expansions, we can clean them up a bit as follows\(^9\)

(%i5) f1 : v0*t*cos(theta)$
(%i6) f2 : expand(xb/f1)$
(%i7) xb : f1*f2;
(%o7) $v_0 t \cos \theta \left(1 - \frac{bt}{2m}\right)$

(%i8) f1 : part(yb,2) + part(yb,4)$
(%i9) f2 : factor(yb - f1)$
(%i10) yb : f1 + f2;
(%o10) $v_0 t \sin \theta - \frac{1}{2}gt^2 - \frac{bt^2}{6m} \left(3v_0 \sin \theta - gt\right)$

We notice first—reassuringly—that the terms that are zeroth-order in $b$, i.e., $x = v_0 t \cos \theta$ and $y = v_0 t \sin \theta - \frac{1}{2}gt^2$, are exactly what we would expect for the projectile in the absence of air resistance. More generally, to find the range when $b$ is small, we begin by removing the root at $t = 0$ and solving for the roots of the Taylor expansion of $y$ with the statements

\(^9\)Note that, in the version of MAXIMA you are using, the order of the terms in `%o4` may be different from that shown, in which case the extraction of the proper parts—those independent of $b$—in the statement in `%i8` may require different numbers.
14.3. FINDING ROOTS SYMBOLICALLY WITH MAXIMA 317

(%i11) expand( yb/t );

(%o11) \frac{v_0 \sin \theta - \frac{1}{2} gt - \frac{bt}{6m} (3v_0 \sin \theta - gt)}{}

(%i12) solve( % = 0, t );

(%o12) \{[t = \text{mess}, t = \text{mess}]\}

To sort out which of these two roots is the physically significant one, we evaluate the limit of this expression as \( b \to 0 \) with the statements

(%i13) limit( %, b, 0 );

Is \( g \) positive or negative? \( p \);

Is \( m \) positive or negative? \( p \);

(%o13) \left\{ t = \frac{2v_0 \sin \theta}{g}, t = \text{infinity} \right\}

Clearly, only the first root is of interest, and we extract it from the more general expression %o12 with the statement

(%i14) part( %o12, 1, 2 );

(%o14) \frac{3m}{2b} + \frac{3v_0}{2g} \sin \theta - \frac{\sqrt{3} \sqrt{3v_0^2b^2 \sin^2 \theta - 2v_0bgm \sin \theta + 3g^2m^2}}{2bg}

Finally, for consistency with approximations already made, we find this root to first order in \( b \) with the statement

(%i15) time : expand( taylor( %, b, 0, 1 ) );

(%o15) \frac{2v_0 \sin \theta \cos \theta}{g} - \frac{2v_0^2 \sin^2 \theta}{3g^2m} b

and find the range by substituting this result into the first order expression for the horizontal coordinate with the statements

(%i16) range : xb, t=time;

(%o16) \text{mess}

In examining the mess, however, we note that some terms in \( b^2 \) have crept in. The statement

(%i17) expand( taylor( range, b, 0, 1 ) );

(%o17) \frac{2v_0^2 \cos \theta \sin \theta}{g} - \frac{8bv_0^3 \cos \theta \sin^2 \theta}{3g^2m}

removes those terms and simplifies the mess somewhat. Now, however, we note the common factor \( 2v_0^2 \sin \theta \cos \theta/g \) in both terms, so the statements

(%i18) f1 : 2*v0^2*sin(theta)*cos(theta)/g$

(%i19) f2 : expand(%o17/f1)$

(%i20) range : f1*f2;

(%o20) \frac{v_0^2}{g} (2 \sin \theta \cos \theta) \left(1 - \frac{4v_0b}{3mg} \sin \theta \right) + \ldots

yields a simple final result. Note in particular that, in the absence of viscous resistance (\( b = 0 \)), this result reduces to the known result \( (v_0^2/g) \sin 2\theta \).
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.

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
   
   $$x_{\text{mid}} = \frac{1}{2} (x_{\text{min}} + x_{\text{max}}) \quad \text{and} \quad f_{\text{mid}} = f(x_{\text{mid}})$$  \hspace{1cm} (14.34)

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.

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 14.8: 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., \( df(x)/dx \big|_{x_n} = f_n' \).
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

\[
\frac{\Delta f}{\Delta x} = \frac{0 - f_n}{x_{n+1} - x_n} = f_n' \quad \implies \quad x_{n+1} = x_n - \frac{f_n}{f_n'} \tag{14.35}
\]

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

\[
f(x) = \prod_i (x - x_i)
\]  

(14.36)

(2) finds a couple of relationships satisfied by polynomials in this form, (3) assumes with only weak justification \textit{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

\[
x_{n+1} = x_n - \frac{N}{G \pm \sqrt{(N - 1)(NH - G^2)}}
\]  

(14.37)

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.\(^1\)\(^\text{10} \) 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 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

\(^{10}\) 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.
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.

### 14.8 Finding Roots Numerically with MATLAB

> Note: All MATLAB program (.m) files referred to in this chapter are available in the directory $\$HEAD/matlab, 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 MATLAB’s default search path. If so, the files can be found by MATLAB 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.

Root-finding functions in MATLAB include

- **roots**, which finds 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.

- **fzero**, seeks a real root of a nonlinear, continuous function of a single variable. It is invoked with a statement like

  \[
  q = \text{fzero}( @\text{func}, x )
  \]

  in the newer—and now preferred—syntax or the general form

  \[
  q = \text{fzero}( '\text{func}', x )
  \]

  in the older syntax. In either form, \( \text{func} \) 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. \( x \) 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, and \( q \) stores the returned root. If \( x \) is a scalar, fzero begins by searching for an interval around \( x \) within which \( \text{func} \) changes sign. Then, taking that interval (or the one specified in a two-component value for \( x \)) as a start, 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). To change that tolerance, we must reset the property TolX by invoking a statement like
>> opt = optimset( 'TolX', 1.0e-10 )

to create a variable opt containing the desired value of TolX and then invoke fzero with the
expanded statement

q = fzero( @func, x, opt ) or q = fzero( 'func', x, opt )

which replaces the default value of TolX with the newly specified one but keeps all other
options at their default values. Further, should the function whose roots are sought contain
parameters, their values can be provided as additional arguments after the argument opt.11

In addition, Section 14.B contains a listing of the Lawrence-written file lubisect.m, which
defines a MATLAB function that uses the method of bisection. It is called with a statement of the
form

q = lubisect( @func, xlb, xub, tol, itmax )

or

q = lubisect( 'func', xlb, xub, tol, itmax )

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.

14.8.1 Classical Turning Points

We might, for example, find the turning points for the potential energy in Eq. (14.3) with the
MATLAB statements

>> 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.

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 which we store in the current
directory with the name turnpts.m.12 Then, we would find the roots of the defined polynomial
with the MATLAB statements13

11If parameters are to be supplied but all default options are to be accepted, use [] as a place holder in the position
where opt would otherwise be stated.
12See Section 3.18.2 for ways to set the current directory.
13In more recent versions of MATLAB, fzero does not provide the information about the interval in which the root
is found.
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;
end

>> q = fzero( @turnpts, -50.0 )   \% Find lowest root.
q = -48.2683
>> q = fzero( @turnpts, -10.0 )  \% Find middle root.
q = -11.0804
>> q = fzero( @turnpts, 10.0 )   \% Find highest root.
Zero found in the interval: [9.2, 10.5657].
q = 9.3487

For this method, we needed some information about the location of each root in order to start the process, and we based our input on the information extracted from Fig. 14.1.

The same function M-file can alternatively be used in an invocation of lubisect with the statement

>> q = lubisect( @turnpts, -50.0, 0.0, 1.0e-10, 0.0, 47.5, 1.0e-10, 50 )
Number of iterations = 35
q = -48.2683

With different starting values, the routine lubisect would, of course, converge on the remaining roots.

14.8.2 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 kvals in our statements to MATLAB. 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 $0 \leq \kappa \leq 4.0$, 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

>> kvals = transpose(0.0:0.1:4.0);
>> 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 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
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.

```matlab
>> q = []; >> for i = 1:41 q = [ q roots( coeffs(i,:) ) ]; end; >> q = transpose(q);
```

The result is a 2 column × 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', 14); 
>> grid on; 
>> xlabel( 'kappa', 'FontSize', 14 ); 
>> ylabel( 'omega', 'FontSize', 14 ); 
>> hold off 
```

The resulting graph is shown in Fig. 14.9.

### 14.8.3 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
Table 14.2: The MATLAB function projectile.m.

```matlab
function fct = projectile( tau, alpha, theta )
% PROJECTILE: Evaluates function involved in projectile motion.
% The function projectile returns the value of a dimensionless
% function that plays a role in determining the range of
% a projectile fired at an elevation theta into a medium
% whose viscosity is characterized by the dimensionless
% parameter alpha.
tmp = 1.0+alpha*sin(theta); % Calculate coefficient
fct = tau - tmp*(1-exp(-tau)); % Return value of f(tau)
end
```

A (dimensionless) viscous damping coefficient $\alpha = \frac{bv_0}{mg}$ 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 current directory in a file named `projectile.m`. For example, the upper left frame in Fig. 14.4 could have been created by setting several preliminary variables with the statements

```matlab
>> thetavals = pi * [0.0:6.0] / 12.0; % Set $\theta = 0^\circ, 15^\circ, \ldots, 90^\circ$ in radians.
>> tau=[0.0:0.01:2.0]; % Set values of $\tau$.
and then executing the statements
```

```matlab
>> alpha = 0.1; % Choose $\alpha$.
>> theta = thetavals(1); % Plot graph for $\theta = 0^\circ$.
>> plot( tau, projectile(tau, alpha, theta ), ... % 'Color', 'black', 'LineWidth', 3.0 )
      'Color', 'black', 'LineWidth', 3.0 )
>> set( gca, 'Xlim', [0.0,0.25], 'Ylim', ... % Set scaling, size of labels.
      [-0.005,0.005], 'FontSize', 14 )
>> hold on;
>> for i=2:7 % Overplot graphs for $\theta = 15^\circ, \ldots, 90^\circ$.
    theta = thetavals(i);
    plot( tau, projectile( tau, alpha, theta ), ... % 'Color', 'black', 'LineWidth', 3.0 )
        'Color', 'black', 'LineWidth', 3.0 );
  end
>> grid on;
```

The remaining frames in Fig. 14.4 were produced with similar statements.

For this example, we choose MATLAB's built-in routine `fzero`\textsuperscript{14}. Starting with a fresh MATLAB 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\textsuperscript{15}

\textsuperscript{14}The function of interest is not a polynomial, so `roots` is not an option. The routine `fliptight` would work, but it requires two initial guesses. Since we want to find many roots, our task will be simplest if we are obliged to provide only one initial guess.

\textsuperscript{15}In some versions of MATLAB, the invocation of `fzero` in this coding will produce a warning message, but the indicated problem appears not to create an error in the output. The display of the message can be suppressed with the statement `warning off`. 
Set \( \alpha \) and \( \theta \).

Set initial guess for root.

Invoke \texttt{fzero}, displaying result.

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

\[
\text{range} = \cos(\theta) \times \left( 1 - \exp(-q) \right) / \alpha
\]

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, say increments of 1\(^\circ\) between consecutive values of \( \theta \). \textit{Continuing in the same session with MATLAB}, we prepare for a calculational loop with the statements

\[
\text{thetadeg} = 0.0:1.0:90.0;
\]

\[
\text{thetarad} = \text{pi} \times \text{thetadeg} / 180.0;
\]

Then, choosing \( \alpha = 0.1 \) and noting from the first frame in Fig. 14.4 that all roots lie below \( \tau = 0.2 \), we continue with the statements\(^{16}\)

\[
\text{alpha} = 0.1; \text{tau0} = 0.2;
\]

\[
\text{for } i = 1:91
\]

\[
\text{theta} = \text{thetarad}(i);
\]

\[
\text{roots}(i) = \text{fzero}( \text{@projectile}, \text{tau0}, [], \alpha, \text{theta} );
\]

\[
\text{end}
\]

\[
\text{range1} = \cos(\text{thetarad}). \times (1 - \exp(-\text{roots}))/\alpha;
\]

\[
\text{plot}( \text{thetadeg}, \text{range1}, \text{Color}, \text{black'}, ... \text{LineWidth}, 3.0 );
\]

\[
\text{grid on};
\]

\[
\text{set}( \text{gca}, \text{FontSize}, 14 );
\]

\[
\text{xlabel}(\theta(\text{deg}'), \text{FontSize}, 14);
\]

\[
\text{ylabel}(R(\theta'), \text{FontSize}, 14);
\]

Repeating this process with \((\alpha = 0.2, \tau_0 = 0.4)\), with \((\alpha = 0.4, \tau_0 = 0.8)\), and with \((\alpha = 0.8, \tau_0 = 1.5)\), storing the results in \texttt{range2}, \texttt{range4}, and \texttt{range8}, respectively, and overplotting each on the axes created above, we finally produce Fig. 14.10. As \( \alpha \) increases \((b \text{ increases}, v_0 \text{ increases}, m \text{ decreases, or some combination})\), 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 \texttt{range} to find the maximum range and the approximate firing angle to achieve it in each case. For example, the statements

\[\text{Several warnings like the one identified in the previous footnote may appear as the loop in this coding is executed, but they appear not to signal actual errors in the output. In addition, because the roots for the lowest few values of } \theta \text{ are difficult to find, MATLAB returns NaN (not a number) for the values.}\]
Figure 14.10: Range versus $\theta$ for $\alpha = 0.1$ (highest graph), $\alpha = 0.2$, $\alpha = 0.4$, and $\alpha = 0.8$ (lowest graph).

> format long
> max(range1)
ans = 0.91379225626474
> [ range1(44) range1(45) range1(46) ]
ans = 0.913563415463738 0.91379225626474 0.912932244294674

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 $44^\circ$\textsuperscript{17}. Fitting a parabola to the three points and finding its maximum yields the angle $43.71^\circ$\textsuperscript{18}. Similar statements applied to the other three values of $\alpha$ yield that the maximum range and firing angle are $0.841218$ and about $43^\circ$ ($42.53^\circ$) for $\alpha = 0.2$, $0.725949$ and about $40^\circ$ ($40.49^\circ$) for $\alpha = 0.4$, and $0.569568$ and about $37^\circ$ ($37.01^\circ$) for $\alpha = 0.8$, respectively.

### 14.8.4 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 qmupper.m and 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 MATLAB statements

```matlab
>> upper(1) = lubisect( @qmupper, -0.5, 0.5, 1.0e-8, 40 );
>> upper(2) = lubisect( @qmupper, 2.5, 3.5, 1.0e-8, 40 );
>> upper(8) = lubisect( @qmupper, 20.5, 21.5, 1.0e-8, 40 );
```

\textsuperscript{17}Remember that $\theta = 0^\circ$ was stored in element 1 of thetadeg, so element 45 corresponds to $\theta = 44^\circ$.

\textsuperscript{18}See an exercise towards the end the chapter(s) on specific symbol manipulating programs.
Table 14.3: The MATLAB functions qmupper.m and qmlower.m.

function fct = qmupper( s )
% QMUPPER: Returns the value of \( \sin(s) - 0.04\,s \)
fct = sin(s) - 0.04*s;
end

function fct = qmlower( s )
% QMLOWER: Returns the value of \( \sin(s) + 0.04\,s \)
fct = sin(s) + 0.04*s;
end

>> 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 statements

>> upper
>> lower

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 MATLAB are\(^{19}\)

```matlab
>> upper.*cos(upper)./sin(upper) + sqrt(25.0^2-upper.^2)
ans = NaN 0.0000 48.2544 0.0000 42.5628 0.0000 30.6267 0.0000
>> lower.*cos(lower)./sin(lower) + sqrt(25.0^2-lower.^2)
ans = NaN 49.5697 -0.0000 45.9736 0.0000 37.6949 0.0000 18.7393 -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 \( \lim_{s \to 0} s \cot s = 1 \) while \( -\sqrt{25^2-s^2} = -25 \) when \( s = 0 \), these roots don’t satisfy the original equation. Thus, we can assemble a single vector containing the physically meaningful roots with the MATLAB statements

```matlab
>> s = lower(2:9);
>> for i = 1:2:7 s(i)=upper(i+1); end
```

```
```

Finally, remembering that we are interested not so much in the values of \( s \) but in the associated energies, we exploit the equation \( E/V_0 = -\left(1-s^2/c^2\right) \) to find the allowed energies with the statements

```
19The value NaN for the first root in each category arises because of a division by zero in evaluating the function.
```
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.

```
>> energy = -( 1 - s.^2./25.0^2 )
energy = -0.9854 -0.9416 -0.8688 -0.7672 -0.6370 -0.4791 -0.2948 -0.0888
```

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', [0.0,1.5] );
>> plot( [0.0,0.0], [-1.5,0.5], 'Color', 'black' );
>> text( 0.05, 0.4, 'E/V0', 'FontSize', 16 );
>> plot( [0.0,0.1], [0.0,0.0], 'Color', 'black' );
>> text( -0.15, -1.0, '-1.0', 'FontSize', 14 );
>> text( -0.15, 0.0, ' 0.0', 'FontSize', 14 );
```

will produce the energy level diagram of Fig. 14.11. The first of these statements opens the plot window and draws the first energy level, the second freezes the plot window, the third adds the remaining levels, and the fourth adjusts the scaling. The remaining statements add labels and an outline of the well.

### 14.11 Finding Roots Numerically with MAXIMA

MAXIMA makes available several functions for finding roots numerically. For polynomials, the available functions include
• \texttt{nroots( f, a, b )}, which returns the number of real roots of the polynomial \( f \) lying in the range \( a < x \leq b \).

• \texttt{realroots( f, eps )}, which returns values for all of the real roots possessed by the polynomial \( f \), calculating each to an absolute precision specified by \( \text{eps} \) (default in \texttt{rootsepsilon}, which will be \( 10^{-7} \) unless explicitly changed), and

• \texttt{allroots( f )}, which returns single-precision values for all of the roots (real and complex) possessed by the polynomial \( f \).

• \texttt{bfallroots( f )}, which returns bigfloat values for all of the roots (real and complex) possessed by the polynomial \( f \).

Note that \texttt{nroots} and \texttt{realroots} can deal properly with multiple roots and that \texttt{realroots} sets the system variable \texttt{multiplicities} to a list of the multiplicities of each root found, but that \texttt{allroots} may have difficulties dealing with multiple roots. Further details can be found in the MAXIMA manuals.

MAXIMA also contains functions for finding the roots of more general univariate functions, which—of course—includes polynomials as a subclass. We mention in particular the three functions

• \texttt{find_root( expr, x, a, b, [abserr, relerr] )}, which, with \( expr \) an expression depending on \( x \), returns a root of \( expr = 0 \) in the closed interval \([a,b]\) with absolute error \( \text{abserr} \) (default in \texttt{find_root_abs}, which will be zero unless explicitly changed) and relative error \( \text{relerr} \) (default in \texttt{find_root_rel}, which will be zero unless explicitly changed). Here, \( \text{abserr} \) and \( \text{relerr} \) can be presented in either order and are entered in the form \( \text{abserr} = \text{value} \) and \( \text{relerr} = \text{value} \). This function uses the binary search method but will switch to linear interpolation if it determines that the function is smooth enough. The square brackets simply indicate that these quantities are optional, not that they are values in a list.

• \texttt{bf_find_root( expr, x, a, b, [abserr,relerr] )}, which is the same as \texttt{find_root} but returns a bigfloat value.

• \texttt{newton( expr, x, x_0, eps )}, which, with \( expr \) an expression depending on \( x \), returns a root of \( expr = 0 \) by Newton’s method, iterating from the starting point provided by \( x_0 \) and continuing until \( |expr| < \text{eps} \). There is no default value for \( \text{eps} \). The package \texttt{newton1} must be loaded to make this command available.

Simply for peace of mind, let us test the first set of functions with a polynomial that has a multiple root. We execute the statements

\begin{verbatim}
(%i1) pl : expand( (x-1)^2*(x+3)*(x-4) )$  
(%i2) nroots( pl, -10, 10);  
(%o2) 4  
(%i3) realroots( pl, 0.001 );  
(%o3) [x = 1, x = -3, x = 4]  
(%i4) multiplicities;  
(%o4) [ 2, 1, 1 ]
\end{verbatim}

Create polynomial with multiple roots.  
Test \texttt{nroots}.  
Find roots with \texttt{realroots}.  
Display multiplicities.

Clearly, \texttt{nroots} counted all four roots of the given polynomial even though one of the three distinct roots is a double root, while \texttt{realroots} reports only the distinct roots. We must examine the system variable \texttt{multiplicities} to find that one of them is a double root. As illustrated, however, in the statements

\begin{verbatim}
(%i5) allroots(pl);  
(%o5) [x = 1.0, x = 1.0, x = -3.0, x = 4.0]  
(%i6) bfallroots(pl);  
(%o6) [x = -3.0b0, x = 1.0b0, x = 1.0b0, x = 4.0b0]
\end{verbatim}

Find roots with \texttt{allroots}.  
Find roots with \texttt{bfallroots}.
both \texttt{allroots} and \texttt{bfallroots}, on the other hand, report the double root twice. Evidently, this specific (double-rooted) polynomial is not one with which \texttt{allroots} or \texttt{bfallroots} has trouble.

Finally, introducing a non-polynomial function $xe^{-x^2} - x^2 + 10$, for which a graph reveals a root in the interval $2 < x < 4$, we test the second set of functions with the statements

\begin{verbatim}
(%i7) expr : x*exp(-x^2) - x^2 + 10$
(%i8) find_root( expr, x, 2.0, 4.0 );

(%o8) 3.1623003569056242

(%i9) expr, x = %;

(%o9) 1.7604... \times 10^{-16}

(%i10) bf_find_root( expr, x, 2.0, 4.0 );

(%o10) 3.1623003569056242

(%i11) expr, x = %;

(%o11) 1.7597...\times 10^{-16}

(%i12) load(newton1)$

(%i13) newton( expr, x, 2.0, 1.0e-7 );

(%o13) 3.16230037067024

(%i14) expr, x = %;

(%o14) -6.99622... \times 10^{-10}
\end{verbatim}

As implied by the verifications on \%o9 and \%o11, the roots on \%o8 and \%o10 satisfy $xe^{-x^2} - x^2 + 10 = 0$ to within a few parts in $10^{16}$, consistent with the default absolute and relative errors of 0.0. The root given by \texttt{newton}, however, satisfies the equation to only about one part in $10^9$, consistent with the specified error in the call to \texttt{newton}.

### 14.11.1 Classical Turning Points

To find the turning points for the potential energy in Eq. (14.3), we begin by defining the polynomial and asking about the number of roots with the statements

\begin{verbatim}
(%i1) V : x^3/10000+x^2/200-x/500-1/2$
(%i2) nroots( V, -60, 20 );

(%o2) 3

(%i3) float( realroots( V, 0.001 ) );

(%o3) [ x = -48.268, x = -11.081, x = 9.349 ]

(%i4) multiplicities;

(%o4) [ 1, 1, 1 ]

(%i5) float( realroots( V ) );

(%o5) [ x = -48.2682675, x = -11.0804366, 
     x = 9.3487042 ]
\end{verbatim}

The three different commands yield consistent though differently formatted roots.

Alternatively, since a polynomial is simply a subclass of a general function, we might seek the turning points by remembering the bounds determined in Eq. (14.4) and invoke the statements

\begin{verbatim}
(%i6) find_root( V, x, -50.00, -47.50 );

(%o6) -48.26826756675163

(%i7) find_root( V, x, -12.50, -10.00 );

(%o7) -11.08043663084842

(%i8) find_root( V, x, 7.50, 10.0 );

(%o8) 9.348704197600052

(%i9) V, x = [ %o6, %o7, %o8 ];

(%o9) [ 0.0, 0.0, -1.11022... \times 10^{-16} ]
\end{verbatim}
In these statements, we accept the default tolerances (both zero), so the iteration stopped when either the root or the function stayed the same (to the precision of double precision calculations) for two successive iterations. The output at %o9 is consistent with an error no larger than the 16-digit precision of double precision calculations.

Newton’s method gives the same results for the roots, though we only need one starting value. The statements

\[
\begin{align*}
\%i10 & \quad \text{load( newton1 )}\$ \\
\%i11 & \quad \text{newton( V, x, -50.0, 1.0e-7 );} \\
\%i11 & \quad x = -48.2682678 \\
\%i12 & \quad \text{newton( V, x, -10.0 );} \\
\%i12 & \quad x = -11.0804366 \\
\%i13 & \quad \text{newton( V, x, 10.0 );} \\
\%i13 & \quad x = 9.3487042
\end{align*}
\]

find the three roots.

14.11.2 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 \( \alpha = \frac{bv_0}{mg} \) 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). Finding that root and then the range for a particular choice of \( \alpha \) and \( \theta \) is quickly accomplished in MAXIMA with the statements

\[
\begin{align*}
\%i1 & \quad \text{range : cos(theta)*(1-exp(-tau))/alpha;} & \quad \text{Define range [Eq. (14.19)].} \\
\%i1 & \quad \text{cos}\theta (1-e^{-\tau})/\alpha \\
\%i2 & \quad \text{fct : tau - (1+alpha*sin(theta))*(1-exp(-tau));} & \quad \text{Define function [Eq. (14.18)].} \\
\%i2 & \quad \tau - (1 + \alpha \sin \theta)(1 - e^{-\tau}) \\
\%i3 & \quad \text{[ alpha : 0.4, theta : float(%pi*30.0/180.0) ];} & \quad \text{Choose \( \alpha \), \( \theta \).} \\
\%i4 & \quad \text{load( newton1 )}\$ \\
\%i5 & \quad \text{newton( ev(fct), tau, 0.5, 1.0e-7 );} & \quad \text{Find root.} \\
\%i5 & \quad 0.3764379972596656 \\
\%i6 & \quad \text{ev( range, tau = % );} & \quad \text{Evaluate range.} \\
\%i6 & \quad 0.6791768095313501
\end{align*}
\]

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 \( \theta \). We can actually accomplish all of these operations in a single loop with the statements\(^{20}\)

\[
\begin{align*}
\%i7 & \quad \text{for i: 0 thru 90 do block} \\
& \quad \text{ ( thetadeg[i] : i, theta : float( %pi*i/180.0 ),} \\
& \quad \quad \text{ tmp : newton( ev(fct), tau, 0.8, 1.0e-7 ),} \\
& \quad \quad \text{ rng[i] : ev( range, tau = tmp ) )}\$
\end{align*}
\]

(Remember that \( \alpha \) was set to 0.4 at %i3.) Then, we produce the graph in Fig. 14.12 with the statements\(^{20}\) MAXIMA’s construction block(...) allows us to group several statements together so they will be seen as a single statement. In that grouping, individual statements are separated by commas.
Figure 14.12: $R(\theta)$ versus $\theta$ for $\alpha = 0.4$.

Further, rounding to seven digits after the decimal point, by noting the three values,

```
(%i11) [ rng[39], rng[40], rng[41] ];
(%o11) [0.7251743, 0.7259486, 0.7258890]
```

we conclude that the maximum range for $\alpha = 0.4$ occurs near $\theta = 40^\circ$ (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 `%i3` and `%i7`–`%i10` would produce a graph corresponding to that other value of the viscous damping.

### 14.11.3 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). We elect the method of bisection, first creating lists of the lower and upper bounds given in Eqs. (14.26) and (14.27). To that end we execute the statements\(^{21}\)

```
(%i11) ulb : [-0.5, 2.5, 6.0, 8.5, 12.5, 14.5, 19.0, 20.5];
(%i12) uub : [ 0.5, 3.5, 7.0, 9.5, 13.5, 15.5, 20.0, 21.5];
```

\(^{21}\)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.
CHAPTER 14. FINDING ROOTS

\[
\%i3 \text{ llb } : \{-0.5, 3.0, 5.5, 9.5, 11.5, 16.0, 17.5, 22.5, 23.5\}
\]

\[
\%i4 \text{ lub } : \{0.5, 4.0, 6.5, 10.5, 12.5, 17.0, 18.5, 23.5, 24.5\}
\]

Then, we find the roots in each category with the simple loops

\[
\%i5 \text{ for } i : 1 \text{ thru } 8 \text{ do}
\text{ rootu}[i] : \text{find_root}(\sin(s) - 0.04s, s, \text{ulb}[i], \text{uub}[i] );
\]

\[
\%i6 \text{ for } i : 1 \text{ thru } 9 \text{ do}
\text{ rootl}[i] : \text{find_root}(\sin(s) + 0.04s, s, \text{llb}[i], \text{lub}[i] );
\]

Then we concatenate the elements into lists and display them with the statements

\[
\%i7 \text{ rtu } : \text{makelist}(\text{rootu}[i], i, 1, 8 );
\]

\[
\%o7 \quad [0.0, 0.0204776, \ldots, 6.54820, \ldots, 9.05418, \ldots, 13.11879, \ldots, 15.06138, \ldots, 19.76108, \ldots, 20.99428, \ldots]
\]

\[
\%i8 \text{ rtl } : \text{makelist}(\text{rootl}[i], i, 1, 9 );
\]

\[
\%o8 \quad [0.0, 0.0327288, \ldots, 6.03920, \ldots, 9.82883, \ldots, 12.06284, \ldots, 16.42478, \ldots, 18.04325, \ldots, 23.17777, \ldots, 23.86449, \ldots]
\]

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. Because \( \cot(0.0) \), which will appear when Eq. (14.23) is evaluated for the first root in rtu and rtl, is undefined, however, we must begin by removing that first root in each list. The appropriate statements to MAXIMA are

\[
\%i9 \text{ rtu : rest(rtu) ;}
\]

\[
\%i10 \text{ rtablest : rtu*cot(rtu)+sqrt(25.0^2-rtu^2) ;}
\]

\[
\%o10 \quad [-1.42108 \times 10^{-14}, 48.25436 \ldots, -3.55271 \times 10^{-15},
42.56276 \ldots, 3.552713 \times 10^{-14}, 30.62674 \ldots, -4.79616 \times 10^{-14}]
\]

\[
\%i11 \text{ rtltest : rtl*cot(rtl)+sqrt(25.0^2-rtl^2) ;}
\]

\[
\%o11 \quad [49.56967 \ldots, 3.55271 \ldots \times 10^{-15}, 45.97364 \ldots, -2.84217 \ldots \times 10^{-14},
37.69490 \ldots, 4.97379 \ldots \times 10^{-14}, 18.73932, 1.95399 \ldots \times 10^{-14}]
\]

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 rtu[2], rtu[4], rtu[6], rtl[1], rtl[3], rtl[5], and rtl[7]. Thus, we assemble a single list containing the physically meaningful roots with the MAXIMA statements

\[
\%i11 \text{ s : rest( rtl, 1 ) ;}
\]

\[
\%i12 \text{ for } i : 1 \text{ thru } 7 \text{ step } 2 \text{ do}
\text{ s[i] : rtu[i] ;}
\]

\[
\%i13 \text{ s ;}
\]

\[
\%o13 \quad [3.0204776 \ldots, 6.0392037 \ldots, 9.0541858 \ldots, 12.0628480 \ldots, 15.0613891 \ldots, 18.0432570 \ldots, 20.9942864 \ldots, 23.8644949 \ldots]
\]

Finally, remembering that we are interested not so much in the values of 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 statements
Figure 14.13: Energy level diagram for the one-dimensional quantum well when \( c = 25.0 \). The eight short horizontal lines in the vertical column show the energies, measured in units of \( V_0 \). Other lines in the diagram show the contour of the well, with energy \( E/V_0 = -1 \) at the bottom and energy \( E/V_0 = 0 \) at the top.

\[
\text{(i14)} \quad \text{energy} : -(1 - s^2/25.0^2);
\]
\[
\text{(o14)} \quad [-0.98540, -0.94164, -0.86883, -0.76718, -0.63705, -0.47911, -0.29478, -0.08878]
\]

where we have rounded the output to five digits after the decimal point. The statements

\[
\text{(i15)} \quad \text{for } i : 1 \text{ thru } 8 \text{ do } \text{en}[i] : \text{energy}[i], \text{energy}[i] ;
\]
\[
\text{(i16)} \quad \text{rng} : [0.2, 0.8] ;
\]

and then plot the energy-level diagram of Fig. 14.13 with the statement

\[
\text{(i17)} \quad \text{plot2d([discrete, rng, en[1]], [discrete, rng, en[2]], [discrete, rng, en[3]],}
\]
\[
\text{[discrete, rng, en[4]], [discrete, rng, en[5]], [discrete, rng, en[6]],}
\]
\[
\text{[discrete, rng, en[7]], [discrete, rng, en[8]],}
\]
\[
\text{[discrete, [-0.5,0.0,0.0,1.0,1.0,1.5], [0.0,0.0,-1.0,-1.0,0.0,0.0]] },
\]
\[
[x, -0.5, 1.5], [y, -1.5, 0.5], [color, black], [style, [lines,3]],
\]
\[
[legend, false], [ylabel, "E/V0"], [xlabel, "x/a"] );
\]

where the last list invoking the plot option \text{discrete} draws the boundaries of the potential well.

### 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. 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.

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.,

\[
\begin{pmatrix}
a_{11} & a_{12} & a_{13} \\
a_{21} & a_{22} & a_{23} \\
a_{31} & a_{32} & a_{33}
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
x_3
\end{pmatrix}
=
\begin{pmatrix}
b_1 \\
b_2 \\
b_3
\end{pmatrix}
\]

(14.38)

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 inhomogeneities, 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 \( \mathbf{x} \) of unknowns, and an \( n \) element vector \( \mathbf{b} \) of inhomogeneities, we would write the equations compactly in the form

\[
A \mathbf{x} = \mathbf{b}
\]

(14.39)

and we would write their solution formally as

\[
\mathbf{x} = A^{-1} \mathbf{b}
\]

(14.40)

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

\[
x_1 = \frac{1}{|A|} \begin{vmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{32} & a_{33} \end{vmatrix} ; \quad x_2 = \frac{1}{|A|} \begin{vmatrix} a_{11} & b_1 & a_{13} \\ a_{21} & b_2 & a_{23} \\ a_{31} & b_3 & a_{33} \end{vmatrix} ; \quad x_3 = \frac{1}{|A|} \begin{vmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ a_{31} & a_{32} & b_3 \end{vmatrix}
\]

(14.41)

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 inhomogeneities. 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

---

22 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 SOLVING SIMULTANEOUS EQUATIONS

(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.\(^{24}\) 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. (14.38), but the schema is readily extended to \(n\) equations. The process of Gaussian elimination entails

- dividing each equation by the coefficient of \(x_1\), obtaining\(^{25}\)

\[
\begin{pmatrix}
1 & a_{12}' & a_{13}' \\
1 & a_{22}' & a_{23}' \\
1 & a_{32}' & a_{33}'
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
x_3
\end{pmatrix} =
\begin{pmatrix}
b_1' \\
b_2' \\
b_3'
\end{pmatrix}
\] (14.42)

- keeping the first equation and replacing the second and third with the result of subtracting the first from each in turn, obtaining

\[
\begin{pmatrix}
1 & a_{12}' & a_{13}' \\
0 & a_{22}'' & a_{23}'' \\
0 & a_{32}'' & a_{33}''
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
x_3
\end{pmatrix} =
\begin{pmatrix}
b_1' \\
b_2'' \\
b_3''
\end{pmatrix}
\] (14.43)

- dividing the second and third equations by the coefficient of \(x_2\), obtaining

\[
\begin{pmatrix}
1 & a_{12}' & a_{13}' \\
0 & 1 & a_{23}''' \\
0 & 1 & a_{33}'''
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
x_3
\end{pmatrix} =
\begin{pmatrix}
b_1' \\
b_2''' \\
b_3'''
\end{pmatrix}
\] (14.44)

- keeping the first and second equations but replacing the third with the result of subtracting the second from the third, obtaining

\[
\begin{pmatrix}
1 & a_{12}' & a_{13}' \\
0 & 1 & a_{23}''' \\
0 & 0 & a_{33}'''
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
x_3
\end{pmatrix} =
\begin{pmatrix}
b_1' \\
b_2''' \\
b_3'''
\end{pmatrix}
\] (14.45)

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

\(^{24}\)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.

\(^{25}\)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.
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. That is, we can write

$$A = LU$$

where

$$L = \begin{pmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{pmatrix} \quad \text{and} \quad U = \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{pmatrix}$$ (14.46)

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

$$Ax = b \quad \implies \quad LUx = b \quad \implies \quad Ly = b \quad \text{where} \quad y = Ux$$ (14.47)

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.\textsuperscript{27}

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.

**MAXIMA:** Within MAXIMA, solving systems of linear equations is accomplished with the commands \texttt{linsolve} and \texttt{solve} (which actually calls \texttt{linsolve} if the equations are linear). These commands will produce symbolic solutions but will also produce numerical results to the extent permitted by any numbers explicitly present in the equations. The commands use direct methods.

**MATLAB:** The simplest route in MATLAB to find a numerical solution to a system of linear equations was illustrated in Section 3.3.4 and invokes either the operator \texttt{\textbackslash}, which gives the solution to the equation $Ax=b$ as $x = A\backslash b$, or the operator \texttt{/}, which gives the solution to the equation $xA = b$ as $x = b/A$. The backslash operator in particular senses several different cases (triangular matrices, symmetric matrices, even underdetermined and overdetermined cases) and adopts one or another method as appropriate to the case that it senses.

### 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.

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

\[ f_1(x_1, x_2, x_3) = 0 \quad ; \quad f_2(x_1, x_2, x_3) = 0 \quad ; \quad f_3(x_1, x_2, x_3) = 0 \]  

(14.50)

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,

\[ x_1 = x_1^{(0)} + \delta x_1 \quad ; \quad x_2 = x_2^{(0)} + \delta x_2 \quad ; \quad x_3 = x_3^{(0)} + \delta x_3 \]  

(14.51)

and demand that the “corrections” satisfy the equations

\[ 0 = f_1(x_1^{(0)} + \delta x_1, x_2^{(0)} + \delta x_2, x_3^{(0)} + \delta x_3) \]
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

\begin{align}
0 &= f_1(0) + \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(0) + \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 \\
0 &= f_3(0) + \frac{\partial f_3}{\partial x_1} \delta x_1 + \frac{\partial f_3}{\partial x_2} \delta x_2 + \frac{\partial f_3}{\partial x_3} \delta x_3
\end{align}

(14.53)

Here, \( f_1, f_2, f_3 \), and all the derivatives are evaluated at \( x_1^{(0)}, x_2^{(0)}, x_3^{(0)} \) 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 \( x_1^{(1)} = x_1^{(0)} + \delta x_1, \ldots \) as a refined approximation to the desired solution and repeat the process to obtain \( x_1^{(2)}, \ldots \), 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.

**MAXIMA:** For symbolic solution of systems of nonlinear—but polynomial—equations, MAXIMA provides the commands \texttt{algsys} and \texttt{solve} (which actually calls \texttt{algsys} if the equations are nonlinear). For numerical solutions, MAXIMA provides the command \texttt{newton}, which can be supplied with a list of equations as input and is able to address systems as well as individual equations.

**MATLAB:** For numerical solution of systems of nonlinear equations, MATLAB makes available the function \texttt{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 \( ax^2 + bx + c \). \textit{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 \textit{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 \(\alpha\).

14.7. 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 \(\alpha = b v_0 / (mg)\), and then (2) evaluating the range from the expression

\[
\frac{R(\theta, \alpha)}{v_0^2 / g} = \frac{\cos \theta(1 - e^{-\tau})}{\alpha} = \frac{\tau \cos \theta}{\alpha(1 + \alpha \sin \theta)}
\]

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

\[
m \frac{d^2 x_1}{dt^2} = -k x_1 + k(x_2 - x_1)
\]
14.17. EXERCISES

\[
\begin{align*}
m d^2 x_2 \quad &= \quad -k(x_2 - x_1) + k(x_3 - x_2) \\
m d^2 x_3 \quad &= \quad -k(x_3 - x_2) - kx_3
\end{align*}
\]

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 \quad \ldots \text{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 \( \sqrt{2} = 1.41421 \). (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

\[
\begin{align*}
L_0(x) &= 1 \\
L_1(x) &= x \\
L_2(x) &= \frac{1}{2}(3x^2 - 1) \\
L_3(x) &= \frac{1}{2}(5x^4 - 15x^2 + 3) \\
L_4(x) &= \frac{1}{8}(35x^6 - 70x^4 + 15x^2 + 3) \\
L_5(x) &= \frac{1}{16}(63x^8 - 315x^6 + 105x^4 - 5)
\end{align*}
\]

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

\[
\tan \alpha = \pm \tanh \alpha
\]

Find the lowest half dozen natural frequencies for this bar. [A detailed discussion of this exercise can be found in N. H. Fletcher and T. D. Rossing, \textit{The Physics of Musical Instruments}, Second Edition (Springer-Verlag, New York, 1991), pp. 57ff.]
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

\[
\cot \alpha = \pm \tanh \alpha
\]


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

\[
\begin{align*}
  j_0(x) &= \frac{\sin x}{x} ; \\
  j_1(x) &= \frac{\sin x}{x^2} - \frac{\cos x}{x} ; \\
  j_2(x) &= \left( \frac{3}{x^3} - \frac{1}{x} \right) \sin x - \frac{3}{x^2} \cos x
\end{align*}
\]

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 \( c \) 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

\[
I(x) = \frac{\sin^2 x}{x^2}
\]

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\(^{28}\) are \( \{2.4048255577, 5.5200781103, 8.6537279129, 11.7915344391, 14.9309177086, 18.0710639679\} \). *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

\[
V(x) = \frac{-V_0 a^2 (a^2 + x^2)}{8a^4 + x^4} \quad \Rightarrow \quad \frac{V(x)}{V_0} = -\frac{(1 + \bar{\pi}^2)}{8 + \bar{\pi}^2}
\]

where \( \bar{\pi} = x/a \). Using at least three different methods and at least two different computational tools, find the coordinates \( \bar{\pi} \) 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 immoveable 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

\[
I(\theta) = \frac{I_0}{\left(\frac{2J_1(\pi d \sin \theta/\lambda)}{\pi d \sin \theta/\lambda}\right)^2}
\]

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

\[
V_{LJ} = 4\epsilon \left[ \frac{\sigma^{12}}{r^{12}} - \frac{\sigma^6}{r^6} \right] \quad \text{or} \quad \frac{V_{LJ}}{\epsilon} = 4 \left[ \frac{1}{\bar{r}^{12}} - \frac{1}{\bar{r}^6} \right]
\]

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 $(x_2, y_2)$. 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](http://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 $\$\text{HEAD}/\text{data/freefall.dat}$ contains 31 lines, the $i$-th of which contains one numerical value—the value of the position $x$, 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 $\$\text{HEAD}/\text{data/freefall.dat}$ 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
14.17. **EXERCISES**

Figure 14.16: Circuits for Exercise 14.35.

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

\[ \frac{\partial R}{\partial a} = 0 ; \quad \frac{\partial R}{\partial b} = 0 ; \quad \frac{\partial R}{\partial c} = 0 \]

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 `$\text{HEAD/data/freefall.dat}$` 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

\[ m \frac{d^2 x}{dt^2} = F(t) , \quad x(0) = 0 , \quad x(t_f) = a \]

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) \( t_f = 1 \text{ s}, \ a = 10 \text{ m}, \ F(t)/m = 8 \text{ N/kg} \) (constant force), and \( n = 10 \);
(b) same as (a) but with \( n = 100 \);
(c) \( t_f = 1 \text{ s}, \ a = 10 \text{ m}, \ F(t)/m = e^{-t} \text{ N/kg} \) (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.,

\[
V_{12} = \frac{1}{4} (V_{11} + V_7 + V_{13} + V_{17}) \quad ; \quad V_{15} = \frac{1}{4} (V_{14} + V_{10} + 0.0 + V_{20})
\]

(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 \( t_i = 0.05(i - 1) \text{ s} \). 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

$$R = \sum_{i=1}^{N} \left( x_i - A e^{-bt_i} \cos(\omega t_i + \phi) + Ct_i + D \right)^2$$

*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.B Listing of MATLAB 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.')
end

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; end
    i = i + 1;
end

fprintf('
Number of iterations = %d
', i);
if i >= itmax fprintf('Convergence not achieved.
'); end

root = xm;
end
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 algebraic equations, 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—might look like Fig. 15.2. For the transverse motion of the string, Newton’s second law $F = ma$ then requires that

$$\rho(x) \Delta x \frac{\partial^2 u}{\partial t^2} = \tau(x + \Delta x, t) \sin \theta(x + \Delta x, t) - \tau(x, t) \sin \theta(x, t) - \rho(x) \Delta x g$$

(15.1)

or, if we divide by $\Delta x$ and let $\Delta x$ approach zero, that

$$\rho(x) \frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x} \left( \tau(x, t) \sin \theta(x, t) \right) - \rho(x) g$$

(15.2)

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

$$\rho(x) \Delta x \frac{\partial^2 v}{\partial t^2} = \tau(x + \Delta x, t) \cos \theta(x + \Delta x, t) - \tau(x, t) \cos \theta(x, t)$$  \hspace{1cm} (15.3)

or that

$$\rho(x) \frac{\partial^2 v}{\partial t^2} = \frac{\partial}{\partial x} \left( \tau(x, t) \cos \theta(x, t) \right)$$  \hspace{1cm} (15.4)

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

$$\sin \theta = \frac{\Delta u}{\Delta s} = \frac{\Delta u}{\sqrt{\Delta u^2 + (\Delta x + \Delta v)^2}} = \frac{\partial u}{\partial x} \left( \frac{\partial u}{\partial x} \right)^2 + \left( 1 + \frac{\partial v}{\partial x} \right)^2 \right)^{1/2}$$  \hspace{1cm} (15.5)

and that

$$\cos \theta = \frac{\Delta x + \Delta v}{\Delta s} = \frac{1 + \frac{\partial v}{\partial x}}{\left( \frac{\partial u}{\partial x} \right)^2 + \left( 1 + \frac{\partial v}{\partial x} \right)^2 \right)^{1/2}$$  \hspace{1cm} (15.6)

Thus, the equations describing the motion of the string are

$$\rho(x) \frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x} \left[ \tau(x, t) \frac{\partial u}{\partial x} \left( \frac{\partial u}{\partial x} \right)^2 + \left( 1 + \frac{\partial v}{\partial x} \right)^2 \right]^{1/2}$$  \hspace{1cm} (15.7)

$$\rho(x) \frac{\partial^2 v}{\partial t^2} = \frac{\partial}{\partial x} \left[ \tau(x, t) \frac{1 + \frac{\partial v}{\partial x}}{\left( \frac{\partial u}{\partial x} \right)^2 + \left( 1 + \frac{\partial v}{\partial x} \right)^2 \right]^{1/2}$$  \hspace{1cm} (15.8)
15.1. SAMPLE PROBLEMS

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
\[ u(x,0), \quad v(x,0), \quad \frac{\partial u}{\partial x}(x,0), \quad \frac{\partial v}{\partial x}(x,0) \] (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
\[ \Delta u \ll \Delta x \quad \Rightarrow \quad \frac{\partial u}{\partial x} \ll 1 \] (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,
\[ \rho \frac{\partial^2 u}{\partial t^2} = \tau \frac{\partial^2 u}{\partial x^2} - \rho g \] (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
  \[ \tau \frac{d^2 u}{dx^2} = \rho g \] (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
  \[ \frac{\rho \partial^2 u}{\partial t^2} = \tau \frac{\partial^2 u}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} \] (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
\[ \frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} = \nabla^2 u \quad \text{or} \quad \left( \nabla^2 \frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} \right) = 0 \] (15.14)
\(^2\)Since \( [\tau] = N = \text{kg.m/s}^2 \) and \( [\rho] = \text{kg/m} \), \( [\tau/\rho] = (\text{kg.m/s}^2)/(\text{kg/m}) = \text{m}^2/\text{s}^2 \) (Here, the symbol \([ \ldots ] \) stands for
the units of \( \ldots \) )
Here

\[ \nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \]  
(15.15)

\[ = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2}{\partial \phi^2} + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2}{\partial \theta^2} \]  
(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 \((x, y, z)\), \((r, \phi, z)\), 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

rate of heat flow in rod towards positive \(x\) through cross section = \(-KA \frac{\partial u}{\partial x}\)  
(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,

heat flow in time \(\Delta t\) into shaded element across surface at \(x\) = \(-KA \frac{\partial u}{\partial x} \bigg|_x \) \(\Delta t\)  
(15.19)

and

heat flow in time \(\Delta t\) into shaded element across surface at \(x + \Delta x\) = \(+KA \frac{\partial u}{\partial x} \bigg|_{x+\Delta x} \) \(\Delta t\)  
(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

$\text{heat necessary to increase temperature of shaded element by amount } \Delta u = (A \Delta x \rho c) \Delta u$ \hspace{1cm} (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

$$\rho c \frac{\partial u}{\partial t} = \frac{\partial}{\partial x} \left( K \frac{\partial u}{\partial x} \right)$$ \hspace{1cm} (15.23)

or, if $K$ happens to be constant, that

$$\frac{\rho c}{K} \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \implies \frac{\partial^2 u}{\partial x^2} = \frac{1}{\alpha^2} \frac{\partial u}{\partial t}$$ \hspace{1cm} (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

$$\frac{\partial}{\partial x} \left( K \frac{\partial u}{\partial x} \right) = 0 \hspace{1cm} \text{or} \hspace{1cm} (K \text{ constant}) \hspace{1cm} \frac{\partial^2 u}{\partial x^2} = 0$$ \hspace{1cm} (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

$$\nabla \cdot \left( K \nabla u \right) = \rho c \frac{\partial u}{\partial t} \hspace{1cm} \text{or} \hspace{1cm} (K \text{ constant}) \hspace{1cm} \nabla^2 u = \frac{1}{\alpha^2} \frac{\partial u}{\partial t}$$ \hspace{1cm} (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

$$\nabla^2 u = 0$$ \hspace{1cm} (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

$$\nabla \times \mathbf{E} = 0 \implies \mathbf{E} = -\nabla V \hspace{1cm} \nabla \cdot \mathbf{D} = \nabla \cdot \varepsilon \mathbf{E} = \rho \implies \nabla \cdot (\varepsilon \nabla V) = -\rho$$ \hspace{1cm} (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

\[ \mathcal{L}\{u_1\} = 0, \, \mathcal{L}\{u_2\} = 0 \quad \implies \quad \mathcal{L}\{au_1 + bu_2\} = a\mathcal{L}\{u_1\} + b\mathcal{L}\{u_2\} = 0 \quad (15.35) \]

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

\[ \frac{\partial^2 u}{\partial x^2} - \frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} = 0 \quad \implies \quad A = 1, B = 0, C = -\frac{1}{c^2} \quad \implies \quad AC - B^2 = -\frac{1}{c^2} < 0 \quad (15.36) \]

and conclude that the wave equation is hyperbolic. For the diffusion equation,

\[ \frac{\partial^2 u}{\partial x^2} - \frac{1}{\alpha^2} \frac{\partial u}{\partial t} = 0 \quad \implies \quad A = 1, B = C = 0 \quad \implies \quad AC - B^2 = 0 \quad (15.37) \]

and the diffusion equation is parabolic. Finally, for the Laplace equation,

\[ \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \quad \implies \quad A = C = 1, B = 0 \quad \implies \quad AC - B^2 = 1 \quad (15.38) \]

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

\[
\text{mass added during interval from } t \text{ to } t + \Delta t = \left[ \rho(t + \Delta t) - \rho(t) \right] \Delta x \Delta y \Delta z \approx \frac{\partial \rho}{\partial t} \Delta t \Delta x \Delta y \Delta z \quad (15.39)
\]

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 of height \( v_x(x) \Delta t \Delta y \Delta z \) 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 over these two surfaces by the fluid flow is given by

\[
\text{net mass transported into volume through sides parallel to } yz \text{ plane in time } t \text{ to } t + \Delta t = -\left( \frac{\partial (\rho v_x)}{\partial x} \right) \Delta t \Delta x \Delta y \Delta z \quad (15.40)
\]

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

\[
\text{net mass transported into volume through all sides in time } t \text{ to } t + \Delta t = -\left( \frac{\partial (\rho v_x)}{\partial x} + \frac{\partial (\rho v_y)}{\partial y} + \frac{\partial (\rho v_z)}{\partial z} \right) \Delta t \Delta x \Delta y \Delta z = -\nabla \cdot (\rho \mathbf{v}) \Delta t \Delta x \Delta y \Delta z \quad (15.41)
\]
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,

\[
\nabla \cdot (\rho v) + \frac{\partial \rho}{\partial t} = 0
\]

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.\(^4\) 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.,

\[
T_{rs} = \begin{pmatrix}
\text{the component in the } s \text{ direction of} \\
\text{the force on a surface of unit area} \\
\text{oriented perpendicular to the } r \text{ direction}
\end{pmatrix}
\]

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)

\(^4\)The 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

$$\mathbf{T}_x = T_{xx} \hat{i} + T_{xy} \hat{j} + T_{xz} \hat{k}$$  \hspace{1cm} (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

$$F_x = \left[ T_{xx}(x + \Delta x, y, z) - T_{xx}(x, y, z) \right] \hat{i} + \left[ T_{xy}(x + \Delta x, y, z) - T_{xy}(x, y, z) \right] \hat{j}$$

$$+ \left[ T_{xz}(x + \Delta x, y, z) - T_{xz}(x, y, z) \right] \hat{k}$$ \hspace{1cm} (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.\footnote{It is extremely easy to become very confused about these signs. Think about them several times.}

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

$$F_y = \left( \frac{\partial T_{yx}}{\partial y} \hat{i} + \frac{\partial T_{yy}}{\partial y} \hat{j} + \frac{\partial T_{yz}}{\partial y} \hat{k} \right) \Delta x \Delta y \Delta z$$  \hspace{1cm} (15.46)

and

$$F_z = \left( \frac{\partial T_{zx}}{\partial z} \hat{i} + \frac{\partial T_{zy}}{\partial z} \hat{j} + \frac{\partial T_{zz}}{\partial z} \hat{k} \right) \Delta x \Delta y \Delta z$$  \hspace{1cm} (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\footnote{For convenience, we define $f_x$ to be a force per unit mass.} $\rho f_x \Delta x \Delta y \Delta z$, we conclude that the $x$ component of the net force on the element in question would be given by

$$F_x = \left( \rho f_x + \frac{\partial T_{xx}}{\partial x} + \frac{\partial T_{yx}}{\partial y} + \frac{\partial T_{xz}}{\partial z} \right) \Delta x \Delta y \Delta z$$  \hspace{1cm} (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

$$\rho a_x = \rho f_x + \frac{\partial T_{xx}}{\partial x} + \frac{\partial T_{yx}}{\partial y} + \frac{\partial T_{xz}}{\partial z}$$  \hspace{1cm} (15.49)

Similar equations could be deduced for the $y$ and $z$ components of the equation of motion for the fluid.
Evaluating the acceleration, however, introduces a subtle complication, since \( a_x \neq \partial v_x / \partial t \). 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

\[
a = \frac{\Delta \mathbf{v}}{\Delta t} = v_x \frac{\partial \mathbf{v}}{\partial x} + v_y \frac{\partial \mathbf{v}}{\partial y} + v_z \frac{\partial \mathbf{v}}{\partial z} + \frac{\partial \mathbf{v}}{\partial t} = \frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla) \mathbf{v}
\]  

which 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

\[
\rho \left( \frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla \right) v_x = \rho f_x + \frac{\partial T_{xx}}{\partial x} + \frac{\partial T_{yx}}{\partial y} + \frac{\partial T_{zx}}{\partial z}
\]

The \( y \) and \( z \) components have similar statements.\(^7\)

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., \( T_{yx} = \mu \partial v_x / \partial y \), where \( \mu \) is the viscosity of the fluid (which in effect is defined by this relationship).\(^8\)

### 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 \( T_{x y} = T_{y x} = \ldots = 0 \). 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.
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

\[
\frac{\partial}{\partial x} (\rho v_x) = -\frac{\partial \rho}{\partial x}; \quad \rho \left( \frac{\partial v_x}{\partial t} + v_x \frac{\partial v_x}{\partial x} \right) = -\frac{\partial p}{\partial x} \tag{15.55}
\]

(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

\[
\frac{\partial}{\partial x} (\rho v_x) = -\frac{\partial \rho}{\partial t}; \quad \rho \frac{\partial v_x}{\partial t} = -K \frac{\partial \rho}{\partial x} \tag{15.56}
\]

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

\[
\frac{\partial}{\partial x} (\rho v_x) = -\frac{\partial \rho}{\partial t}; \quad \rho \frac{\partial v_x}{\partial t} = -K_0 \frac{\partial \rho}{\partial x} \tag{15.58}
\]

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

\[
\rho = \rho_0 + \delta \rho(x,t); \quad v_x = \delta v_x(x,t) \tag{15.59}
\]

Then, ignoring all but the linear terms in small quantities, we find that the equations now are

\[
\rho_0 \frac{\partial \delta v_x}{\partial x} = -\frac{\partial \delta \rho}{\partial t}; \quad \rho_0 \frac{\partial \delta v_x}{\partial t} = -K_0 \frac{\partial \delta \rho}{\partial x} \tag{15.60}
\]

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

\[
\rho_0 \frac{\partial^2 \delta v_x}{\partial t \partial x} = \frac{\partial^2 \delta \rho}{\partial t^2}; \quad \rho_0 \frac{\partial^2 \delta v_x}{\partial x \partial t} = -K_0 \frac{\partial^2 \delta \rho}{\partial x^2} \tag{15.61}
\]

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 \( p = nRT/V = MRT/(Vn_m) \), 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

\[
p \left( \frac{M}{\rho} \right)^\gamma = p_0V_0^\gamma \quad \implies \quad p = \frac{p_0V_0^\gamma}{M^\gamma \rho^\gamma} \quad \implies \quad K = \frac{dp}{d\rho} = \gamma\frac{p_0V_0^\gamma}{M^\gamma \rho^{\gamma-1}} \tag{15.63}
\]
15.1. SAMPLE PROBLEMS

or, evaluating at \( \rho = \rho_0 \) and noting the relationship \( V_0 = M/\rho_0 \), that

\[
K_0 = \frac{dp}{d\rho_0} = \gamma \rho_0 V_0^2 \rho_0^{-1} = \gamma \frac{\rho_0 M^\gamma}{M^\gamma \rho_0} \rho_0^{-1} = \gamma \frac{\rho_0}{\rho_0} = \gamma \frac{RT_0}{m_n}
\] (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 \( R = 8.31 \text{ N m}/(\text{mol K}) \) and—for air—\( m_n = 0.0288 \text{ kg} \) and \( \gamma = 7/5 = 1.4 \), we find at room temperature (70 \( ^\circ \text{F} = 21 \text{ \( ^\circ \text{C} \) = 294 K}) that

\[
v = \sqrt{\gamma RT_0} = \sqrt{1.4 \times 8.31 \times 294 / 0.0288} \text{ m/s} = 344 \text{ m/s} \] (15.65)

A further interesting connection appears if we remember from the equipartition theorem of kinetic theory that the rms speed \( v_{\text{rms}} \) of molecules of mass \( m \) in a gas at absolute temperature \( T_0 \) satisfies \( \frac{1}{2} m v_{\text{rms}}^2 = \frac{3}{2} kT_0 \) or, on multiplying by Avogadro’s number, \( \frac{1}{2} m_n v_{\text{rms}}^2 = \frac{3}{2} RT_0 \). We conclude that

\[
v_{\text{rms}} = \sqrt{\frac{3RT_0}{m_n}} \implies v = \sqrt{\frac{\gamma RT_0}{m_n}} = \sqrt{\frac{\gamma 3RT_0}{3 m_n}} = \sqrt{\frac{\gamma}{3}} v_{\text{rms}} \] (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

\[
-\frac{d}{dx} \left( \alpha(x) \frac{d\varphi(x)}{dx} \right) + \beta(x) \varphi(x) = f(x) \quad \text{or} \quad -\alpha(x) \frac{d^2\varphi(x)}{dx^2} - \alpha'(x) \frac{d\varphi(x)}{dx} + \beta(x) \varphi(x) = f(x)
\] (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

\[
\varphi(0) = p \] (15.68)

and

\[
\left[ \alpha(x) \frac{d\varphi(x)}{dx} + \gamma \varphi(x) \right]_{x=L} = q \] (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

\[-\frac{\partial}{\partial x} \left( \alpha_x(x,y) \frac{\partial \varphi}{\partial x} \right) - \frac{\partial}{\partial y} \left( \alpha_y(x,y) \frac{\partial \varphi}{\partial y} \right) + \beta(x,y) \varphi = f(x,y) \] (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

\[ \left( \alpha_x \frac{\partial \varphi}{\partial x} \mathbf{i} + \alpha_y \frac{\partial \varphi}{\partial y} \mathbf{j} \right) \cdot \hat{n} = q \quad \text{(on } \Gamma_2) \] (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

\[-\alpha_i \frac{d^2 \varphi(x)}{dx^2} \bigg|_{x_i} - \alpha_i' \frac{d\varphi(x)}{dx} \bigg|_{x_i} + \beta_i \varphi_i = f_i \] (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

\[ \frac{d\varphi(x)}{dx} \bigg|_{x_i} \approx \frac{\varphi_{i+1} - \varphi_i}{\Delta x} \quad \text{or} \quad \frac{d\varphi(x)}{dx} \bigg|_{x_i} \approx \frac{\varphi_i - \varphi_{i-1}}{\Delta x} \quad \text{or} \quad \frac{d\varphi(x)}{dx} \bigg|_{x_i} \approx \frac{\varphi_{i+1} - 2\varphi_i + \varphi_{i-1}}{2\Delta x} \] (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

\[ \frac{d^2 \varphi(x)}{dx^2} \bigg|_{x_i} \approx \frac{\varphi(x)}{dx} \bigg|_{x_i} \approx \frac{\varphi_{i+1} - 2\varphi_i + \varphi_{i-1}}{\Delta x^2} \] (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. 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

$$-\alpha_i \left( \frac{\varphi_{i+1} - 2 \varphi_i + \varphi_{i-1}}{\Delta x^2} \right) - \alpha_i' \left( \frac{\varphi_{i+1} - \varphi_{i-1}}{2 \Delta x} \right) + \beta_i \varphi_i = f_i$$

or, on multiplying by $\Delta x^2$ and collecting terms with the same index on $\varphi$, that

$$\left( -\alpha_i + \frac{\alpha_i' \Delta x}{2} \right) \varphi_{i-1} + \left( 2 \alpha_i + \beta_i \Delta x^2 \right) \varphi_i + \left( -\alpha_i - \frac{\alpha_i' \Delta x}{2} \right) \varphi_{i+1} = f_i \Delta x^2$$

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 **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

$$\varphi_0 = p$$

For the **mixed** boundary condition of Eq. (15.69), we use a central difference approximation to the derivative in the boundary condition, finding that

$$\alpha_N \left( \frac{\varphi_{N+1} - \varphi_{N-1}}{2 \Delta x} \right) + \gamma \varphi_N = q \quad \implies \quad \varphi_{N+1} = \varphi_{N-1} + \frac{2 \Delta x}{\alpha_N} (q - \gamma \varphi_N)$$

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

$$-2 \alpha_N \varphi_{N-1} + \left[ 2 \left( \alpha_N + \gamma \Delta x \right) + \left( \beta_N + \frac{\gamma \alpha_N'}{\alpha_N} \right) \Delta x^2 \right] \varphi_N = \left( f_N + \frac{\alpha_N' q}{\alpha_N} \right) \Delta x^2 + 2q \Delta x$$

With these resolutions of the first and last equations, we arrive at the now fully defined and complete set

$$\begin{align*}
\varphi_0 &= p \\
\left( -\alpha_i + \frac{\alpha_i' \Delta x}{2} \right) \varphi_{i-1} + \left( 2 \alpha_i + \beta_i \Delta x^2 \right) \varphi_i + \left( -\alpha_i - \frac{\alpha_i' \Delta x}{2} \right) \varphi_{i+1} &= f_i \Delta x^2, \quad 1 \leq i \leq N-1 \\
-2 \alpha_N \varphi_{N-1} + \left[ 2 \left( \alpha_N + \gamma \Delta x \right) + \left( \beta_N + \frac{\gamma \alpha_N'}{\alpha_N} \right) \Delta x^2 \right] \varphi_N &= \left( f_N + \frac{\alpha_N' q}{\alpha_N} \right) \Delta x^2 + 2q \Delta x
\end{align*}$$

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 **Dirichlet** boundary conditions at $x = 0$ and **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

\[
\begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
? & ? & ? & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & ? & ? & ? & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & ? & ? & ? & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & ? & ? & ? & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & ? & ? & ? & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & ? & ? & ? & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & ? & ? & ? & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & ? & ? & ? \\
\end{pmatrix}
\begin{pmatrix}
\varphi_0 \\
\varphi_1 \\
\varphi_2 \\
\varphi_3 \\
\varphi_4 \\
\varphi_5 \\
\varphi_6 \\
\varphi_7 \\
\varphi_8 \\
\varphi_9 \\
\varphi_{10}
\end{pmatrix}
= 
\begin{pmatrix}
p \\
p \\
p \\
p \\
p \\
p \\
p \\
p \\
p \\
p \\
\end{pmatrix}
\tag{15.82}
\]

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 determin-
ant of the coefficient matrix is \textit{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 ad-
justed 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.\footnote{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.}

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 \textit{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 \textit{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 \textit{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 \textit{direct} algorithm to solve the discretized equations that convey the
solution, and the third type of error will not be present. When an \textit{iterative} algorithm is used,
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).

## 15.4 Using MATLAB to Solve 1D PDEs via an FDM

### 15.4.1 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 MATLAB statements are

```matlab
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: ' );
```

Next, with the statements

```matlab
dx = L/N; dx2 = dx^2;
x = dx*[ 0 : N ];
```

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 MAT-
LAB’s operator \( \setminus \) 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 \texttt{inhomo}, a vector containing the inhomogeneities. 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
cf(1,1) = 1.0;
cf(N+1,N+1) = cf(N+1,N+1) + 2.0*gamma*dx;
for i=1:N cf(i,i+1) = -alpha; end
cf(1,2) = 0.0;
for i=1:N cf(i+1,i)= -alpha; end
cf(N+1,N) = -2.0*alpha;

to create the necessary matrix and vector, w 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.2, and can be copied from the directory $HEAD/matlab.14

15.4.2 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.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

\[ p = 0.0 \text{ m} \quad ; \quad \gamma = 0.0 \text{ kg/s} \quad ; \quad q = -4.0 \text{ kg.m/s} \]

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 \( 0 \text{ s} \leq t \leq 10 \text{ s} \), so \( L \rightarrow T = 10.0 \text{ s} \). 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 \( 1.0 \text{ m/s} \).

14See the Local Guide for the meaning of $HEAD at your site.
15Note 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

```
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

```
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
```

Once we have entered these parameters, the remaining statements will construct the matrix and generate the solution, storing it in an array named \( \phi \). We will have no further interaction with the program. Presently, the MATLAB 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 MATLAB's main prompt. In particular, we could plot the solution by invoking the statements

```
plot( x, phi, 'Color', 'black', 'LineWidth', 4 )
title('N=20', 'FontSize', 20 )
grid on
```

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

```
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;
```

Then, to plot all solutions in a single display, we execute the statements

```
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);
```
Figure 15.7: Harmonic motion via finite difference analysis. This graph was produced with MATLAB.

```
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 )
```

All of these graphs are shown in Fig. 15.7. 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

```
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
```

which suggest that the solution for \( N = 50 \) is accurate to an absolute error of about ±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 ±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 \textit{initial} velocity and check the \textit{final} velocity with the statements

\[
\text{InitialVelocity} = \frac{\phi(2) - \phi(1)}{(x(2) - x(1))} = -1.3820 \\
\text{FinalVelocity} = \frac{\phi(101) - \phi(100)}{(x(101) - x(100))} = 0.9587
\]

(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 \(-1.3847 \) m/s and \(0.9793 \) m/s, respectively; had we used \( N = 500 \), we would have found those velocities to be \(-1.3855 \) and \(0.9917 \) 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. \textit{Discretizing the domain (preprocessing).} The region of interest is subdivided into a number of small \textit{elements} by appropriately chosen \textit{nodes}, at each of which an approximation to the dependent variable \( \phi \) will be sought. The discretization of the domain through the identification of suitable nodes and elements is known as \textit{preprocessing}.

2. \textit{Selecting the interpolation or shape functions.} The interpolation or shape functions for approximating the dependent variable \textit{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. \textit{Formulating the equations for a single element.} Equations for each element—the \textit{elemental equations}—are formulated using the Ritz variational method, the Galerkin method, or some alternative and less common method.

4. \textit{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. \textit{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. \textit{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 \( x_N = L \).\(^{16}\) 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.\(^{17}\) 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
\[
\tilde{\varphi}^{(e)}(x) = a^{(e)} + b^{(e)} x
\]  
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 \(\tilde{\varphi}^{(e)}_1\) and \(\tilde{\varphi}^{(e)}_2\) 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 \(x^{(e)}_1\) and \(x^{(e)}_2\), we have that
\[
\tilde{\varphi}^{(e)}_1 = a^{(e)} + b^{(e)} x^{(e)}_1 \tag{15.86}
\]
\[
\tilde{\varphi}^{(e)}_2 = a^{(e)} + b^{(e)} x^{(e)}_2
\]
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
\[
\tilde{\varphi}^{(e)}(x) = \sum_{j=1}^{2} N_j^{(e)}(x) \tilde{\varphi}^{(e)}_j \tag{15.87}
\]
where
\[
N_1^{(e)}(x) = \frac{x^{(e)}_2 - x}{l^{(e)}} \tag{15.88}
\]
\[
N_2^{(e)}(x) = \frac{x - x^{(e)}_1}{l^{(e)}} \tag{15.89}
\]
and \(l^{(e)} = x^{(e)}_2 - x^{(e)}_1\) 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
\[
N_1^{(e)}(x^{(e)}_1) = 1 \quad N_1^{(e)}(x^{(e)}_2) = 0 \tag{15.90}
\]
\[
N_2^{(e)}(x^{(e)}_1) = 0 \quad N_2^{(e)}(x^{(e)}_2) = 1
\]
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
where Eq. (15.91) was substituted into Eq. (15.92) for \( r \).

The Galerkin method is, however, applicable even in cases for which a variational formulation cannot be found. In all cases where a variational formulation exists, the Galerkin and Ritz methods are equivalent.

To obtain the approximate solution, the functional is minimized with respect to its variables. In this approach, the minimum of this functional corresponds to the governing differential equation under the given boundary conditions.

The method in which the boundary value problem is formulated in terms of a variational expression or functional. The choice of the weighting functions, which seemed quite arbitrary in the Galerkin approach, is embedded naturally in the development. In Galerkin’s method, since there are that many values of \( \tilde{\varphi} \) 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

\[
R_{i}^{(e)} = \int_{x_{1}^{(e)}}^{x_{2}^{(e)}} W_{i}^{(e)}(x) r(x) \, dx = 0 \quad ; \quad i = 1, 2
\]

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

\[
R_{i}^{(e)} = \int_{x_{1}^{(e)}}^{x_{2}^{(e)}} N_{i}^{(e)} \left[ - \frac{d}{dx} \left( \alpha(x) \frac{d\tilde{\varphi}_{i}^{(e)}}{dx} \right) + \beta(x) \tilde{\varphi}_{i}^{(e)} \right] \, dx - \int_{x_{1}^{(e)}}^{x_{2}^{(e)}} N_{i}^{(e)} f(x) \, dx \quad i = 1, 2
\]

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 \( dv = (d/dx) (\alpha^{(e)} d\tilde{\varphi}_{i}^{(e)}/dx) \, dx \).
Now if Eq. (15.87) is substituted into Eq. (15.94), we find that

\[
R_i^{(e)} = \sum_{j=1}^{2} \varphi_j^{(e)} \int_{x_1^{(e)}}^{x_2^{(e)}} \left( \alpha(x) \frac{dN_i^{(e)}}{dx} \frac{dN_j^{(e)}}{dx} + \beta(x) N_i^{(e)} N_j^{(e)} \right) dx
\]

\[
- \int_{x_1^{(e)}}^{x_2^{(e)}} N_i^{(e)} f(x) dx - \left( \alpha N_i^{(e)} \frac{d\varphi^{(e)}}{dx} \right) \biggr|_{x_1^{(e)}}^{x_2^{(e)}} ; \quad i = 1, 2.
\]

(15.95)

These equations for the weighted residual integrals can also be expressed in matrix form as

\[
\begin{bmatrix}
R_1^{(e)} \\
R_2^{(e)}
\end{bmatrix} =
\begin{bmatrix}
K_{11}^{(e)} & K_{12}^{(e)} \\
K_{21}^{(e)} & K_{22}^{(e)}
\end{bmatrix}
\begin{bmatrix}
\varphi_1^{(e)} \\
\varphi_2^{(e)}
\end{bmatrix} - 
\begin{bmatrix}
b_1^{(e)} \\
b_2^{(e)}
\end{bmatrix} - 
\begin{bmatrix}
g_1^{(e)} \\
g_2^{(e)}
\end{bmatrix}
\]

(15.96)

or more compactly as

\[
\{ R^{(e)} \} = [K^{(e)}] \{ \varphi^{(e)} \} - \{ b^{(e)} \} - \{ g^{(e)} \}
\]

(15.97)

where

\[
K_{ij}^{(e)} = \int_{x_1^{(e)}}^{x_2^{(e)}} \left( \alpha(x) \frac{dN_i^{(e)}}{dx} \frac{dN_j^{(e)}}{dx} + \beta(x) N_i^{(e)} N_j^{(e)} \right) dx
\]

(15.98)

\[
b_i^{(e)} = \int_{x_1^{(e)}}^{x_2^{(e)}} N_i^{(e)} f(x) dx
\]

(15.99)

and

\[
g_i^{(e)} = \left( \alpha N_i^{(e)} \frac{d\varphi^{(e)}}{dx} \right) \biggr|_{x=x_1^{(e)}}^{x=x_2^{(e)}} - \left( \alpha N_i^{(e)} \frac{d\varphi^{(e)}}{dx} \right) \biggr|_{x=x_1^{(e)}}^{x=x_2^{(e)}}
\]

(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

\[
K_{11}^{(e)} = K_{22}^{(e)} = \frac{\alpha^{(e)} l^{(e)}}{3} + \frac{\beta^{(e)} l^{(e)}}{3}
\]

(15.101)

\[
K_{12}^{(e)} = K_{21}^{(e)} = -\frac{\alpha^{(e)} l^{(e)}}{6} + \frac{\beta^{(e)} l^{(e)}}{6}
\]

(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

\[
b_1^{(e)} = b_2^{(e)} = f^{(e)} \frac{l^{(e)}}{2}
\]

(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

$$
g_1^{(e)} = -\alpha (x_1^{(e)}) \frac{d\tilde{\varphi}_1^{(e)}}{dx} \bigg|_{x_1^{(e)}}
$$

$$
g_2^{(e)} = \alpha (x_2^{(e)}) \frac{d\tilde{\varphi}_2^{(e)}}{dx} \bigg|_{x_2^{(e)}}
$$

(15.104)

With $[K]$, $\{b\}$, and $\{g\}$ given by Eqs. (15.101)–(15.104), the elemental equation obtained by requiring $R_i^{(e)}$ as given by Eq. (15.96) to be zero for all $i$ is

$$
\begin{bmatrix}
K_{11}^{(e)} & K_{12}^{(e)} \\
K_{21}^{(e)} & K_{22}^{(e)}
\end{bmatrix}
\begin{bmatrix}
\varphi_1^{(e)} \\
\varphi_2^{(e)}
\end{bmatrix}
= 
\begin{bmatrix}
\varphi_1^{(1)} \\
\varphi_2^{(1)}
\end{bmatrix}
+ 
\begin{bmatrix}
g_1^{(e)} \\
g_2^{(e)}
\end{bmatrix}
$$

(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

$$
\begin{bmatrix}
K_{11}^{(1)} & K_{12}^{(1)} \\
K_{21}^{(1)} & K_{22}^{(1)}
\end{bmatrix}
\begin{bmatrix}
\varphi_1^{(1)} \\
\varphi_2^{(1)}
\end{bmatrix}
= 
\begin{bmatrix}
\varphi_1^{(1)} \\
\varphi_2^{(1)}
\end{bmatrix}
+ 
\begin{bmatrix}
g_1^{(1)} \\
g_2^{(1)}
\end{bmatrix}
$$

(15.106)

which is equivalent to the two equations

$$
K_{11}^{(1)} \varphi_1^{(1)} + K_{12}^{(1)} \varphi_2^{(1)} = b_1^{(1)} + g_1^{(1)}
$$

$$
K_{21}^{(1)} \varphi_1^{(1)} + K_{22}^{(1)} \varphi_2^{(1)} = b_2^{(1)} + g_2^{(1)}
$$

(15.107)

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

$$
K_{11}^{(2)} \varphi_1^{(2)} + K_{12}^{(2)} \varphi_2^{(2)} = b_1^{(2)} + g_1^{(2)}
$$

$$
K_{21}^{(2)} \varphi_1^{(2)} + K_{22}^{(2)} \varphi_2^{(2)} = b_2^{(2)} + g_2^{(2)}
$$

(15.108)

and

$$
K_{11}^{(3)} \varphi_1^{(3)} + K_{12}^{(3)} \varphi_2^{(3)} = b_1^{(3)} + g_1^{(3)}
$$

$$
K_{21}^{(3)} \varphi_1^{(3)} + K_{22}^{(3)} \varphi_2^{(3)} = b_2^{(3)} + g_2^{(3)}
$$

(15.109)

From the relationship between the local and global numbering systems as defined in Eq. (15.84), however, it is clear that $x_2^{(1)} = x_2^{(2)} = x_2$ and $x_2^{(2)} = x_1^{(3)} = x_3$. Imposing the condition of continuity at nodes 2 and 3, we conclude then that $\tilde{\varphi}_2^{(1)} = \tilde{\varphi}_1^{(2)}$ and $\tilde{\varphi}_2^{(2)} = \tilde{\varphi}_1^{(3)}$. Now if we once again invoke the relationship in Eq. (15.84) and recognize the correspondences

$$
\tilde{\varphi}_1^{(1)} \to \tilde{\varphi}_1, \quad \tilde{\varphi}_2^{(1)} = \tilde{\varphi}_1^{(2)} \to \tilde{\varphi}_2, \quad \tilde{\varphi}_2^{(2)} = \tilde{\varphi}_1^{(3)} \to \tilde{\varphi}_3, \quad \text{and} \quad \tilde{\varphi}_2^{(3)} \to \tilde{\varphi}_4
$$
linking the (approximate) values at locally numbered nodes to the (approximate) values \( \tilde{\phi}_1, \tilde{\phi}_2, \tilde{\phi}_3, \) and \( \tilde{\phi}_4 \) at the globally numbered nodes, Eqs. (15.107)–(15.109) become

\[
K_{11}^{(1)} \tilde{\phi}_1 + K_{12}^{(1)} \tilde{\phi}_2 = b_1^{(1)} + g_1^{(1)}
\]
\[
K_{21}^{(i)} \tilde{\phi}_1 + K_{22}^{(i)} \tilde{\phi}_2 = b_2^{(i)} + g_2^{(i)}
\]
\[
K_{11}^{(i)} \tilde{\phi}_2 + K_{12}^{(i)} \tilde{\phi}_3 = b_1^{(i)} + g_1^{(i)}
\]
\[
K_{21}^{(i)} \tilde{\phi}_2 + K_{22}^{(i)} \tilde{\phi}_3 = b_2^{(i)} + g_2^{(i)}
\]
\[
K_{11}^{(i)} \tilde{\phi}_3 + K_{12}^{(i)} \tilde{\phi}_4 = b_1^{(i)} + g_1^{(i)}
\]
\[
K_{21}^{(i)} \tilde{\phi}_3 + K_{22}^{(i)} \tilde{\phi}_4 = b_2^{(i)} + g_2^{(i)}
\]

(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

\[
\begin{bmatrix}
  K_{11}^{(1)} & K_{12}^{(1)} & 0 & 0 \\
  K_{21}^{(1)} & K_{22}^{(1)} + K_{11}^{(1)} & K_{12}^{(2)} & 0 \\
  0 & K_{21}^{(2)} & K_{22}^{(2)} + K_{11}^{(3)} & K_{12}^{(3)} \\
  0 & 0 & K_{21}^{(3)} & K_{22}^{(3)}
\end{bmatrix}
\begin{bmatrix}
  \tilde{\phi}_1 \\
  \tilde{\phi}_2 \\
  \tilde{\phi}_3 \\
  \tilde{\phi}_4
\end{bmatrix}
= 
\begin{bmatrix}
  b_1^{(1)} \\
  b_2^{(1)} + b_1^{(2)} \\
  b_2^{(2)} + b_1^{(3)} \\
  b_2^{(3)} + b_1^{(4)}
\end{bmatrix}
+ 
\begin{bmatrix}
  g_1^{(1)} \\
  g_2^{(1)} + g_1^{(2)} \\
  g_2^{(2)} + g_1^{(3)} \\
  g_2^{(3)} + g_1^{(4)}
\end{bmatrix}
\]

(15.111)

or more compactly

\[
[K][\tilde{\phi}] = \{b\} + \{g\}.
\]

(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

\[
K_{11} = K_{11}^{(1)} = \frac{\alpha^{(1)} f^{(1)}}{l^{(1)}} + \frac{\beta^{(1)} f^{(1)}}{3
\]

\[
K_{ii} = K_{22}^{(i-1)} + K_{11}^{(i)}
= \frac{\alpha^{(i-1)} f^{(i-1)}}{l^{(i-1)}} + \frac{\alpha^{(i)} f^{(i)}}{l^{(i)}} + \frac{\beta^{(i)} f^{(i)}}{3} ; \quad i = 2, 3, 4, \ldots, N - 1
\]

\[
K_{NN} = K_{22}^{(M)} = \frac{\alpha^{(M)} f^{(M)}}{l^{(M)}} + \frac{\beta^{(M)} f^{(M)}}{3
\]

\[
K_{i+1,i} = K_{i,i+1} = K_{12}^{(i)} = -\frac{\alpha^{(i)} f^{(i)}}{l^{(i)}} + \frac{\beta^{(i)} f^{(i)}}{6} ; \quad i = 1, 2, 3, \ldots, N - 1 \quad (15.113)
\]

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

\[
b_1 = b_1^{(1)} = \frac{f^{(1)} l^{(1)}}{2}
\]
\[ b_i = b_2^{(i-1)} + b_1^{(i)} = f^{(i-1)} \frac{l^{(i-1)}}{2} + f^{(i)} \frac{l^{(i)}}{2} \quad i = 2, 3, 4, \ldots, N - 1 \]

\[ b_N = b_1^{(M)} = f^{(M)} \frac{l^{(M)}}{2} \quad (15.114) \]

Finally, we need to evaluate the elements of \( \{ g \} \). Note that all but the first and last entries in \( \{ g \} \) can be written as

\[ g_i = g_2^{(i-1)} + g_1^{(i)} \quad (15.115) \]

If Eq. (15.104) is substituted into Eq. (15.115), we have

\[ g_i = \alpha d\tilde{\phi} \bigg|_{x=x_2^{(i-1)}} - \alpha d\tilde{\phi} \bigg|_{x=x_1^{(i)}} \quad (15.116) \]

However, \( x_2^{(i-1)} = x_1^{(i)} = x_i \) [see Eq. (15.84)] and \( \alpha (d\tilde{\phi}/dx) \) 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

\[ \{ g \} = \begin{bmatrix} - \alpha(x_1) \frac{d\tilde{\phi}}{dx} \big|_{x=x_1} \\ 0 \\ 0 \\ \alpha(x_4) \frac{d\tilde{\phi}}{dx} \big|_{x=x_4} \end{bmatrix} \quad (15.117) \]

In general for a problem with \( M \) elements (\( N \) nodes), we would find that

\[ g_1 = - \alpha(x_1) \frac{d\tilde{\phi}}{dx} \bigg|_{x=x_1=0} \]
\[ g_i = 0 \quad i = 2, 3, 4, \ldots, N - 1 \]
\[ g_N = \alpha(x_N) \frac{d\tilde{\phi}}{dx} \bigg|_{x=x_N=L} \quad (15.118) \]

In consequence of the considerations to this point, our system of equations for \( \{ \tilde{\phi} \} \)—the (approximate) values at the nodes—now assumes the form\(^21\)

\[
\begin{bmatrix}
K_{11} & K_{12} & K_{13} & K_{14} \\
K_{21} & K_{22} & K_{23} & K_{24} \\
K_{31} & K_{32} & K_{33} & K_{34} \\
K_{41} & K_{42} & K_{43} & K_{44}
\end{bmatrix}
\begin{bmatrix}
\tilde{\phi}_1 \\
\tilde{\phi}_2 \\
\tilde{\phi}_3 \\
\tilde{\phi}_4
\end{bmatrix}
= \begin{bmatrix}
b_1 \\
b_2 \\
b_3 \\
b_4
\end{bmatrix}
+ \begin{bmatrix}
g_1 \\
0 \\
0 \\
g_4
\end{bmatrix}
\quad (15.119)
\]

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 \( \{ \tilde{\phi} \} \), we need to incorporate the boundary conditions. First consider the boundary condition of the third kind given in Eq. (15.69). If we replace \( \phi \) by its approximation \( \tilde{\phi} \), solve for \( \alpha (d\tilde{\phi}/dx) \) at \( x = L \), and substitute this into Eq. (15.118), we have that

\[ g_N = \alpha \frac{d\tilde{\phi}}{dx} \bigg|_{x=L} = q - \gamma \tilde{\phi}_N \quad (15.120) \]

\(^{21}\)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
\[ g_4 = \alpha \left. \frac{d\tilde{\psi}}{dx} \right|_{x=L} = q - \gamma \tilde{\psi}_4 \]  
(15.121)

Thus, the last equation in our system becomes
\[ K_{41} \tilde{\psi}_1 + K_{42} \tilde{\psi}_2 + K_{43} \tilde{\psi}_3 + K_{44} \tilde{\psi}_4 = b_4 + q - \gamma \tilde{\psi}_4 \]  
(15.122)
which we can recast as
\[ K_{41} \tilde{\psi}_1 + K_{42} \tilde{\psi}_2 + K_{43} \tilde{\psi}_3 + (K_{44} + \gamma) \tilde{\psi}_4 = b_4 + q \]  
(15.123)

With this rewriting of the fourth equation, our system of equations becomes
\[
\begin{bmatrix}
K_{11} & K_{12} & K_{13} & K_{14} \\
K_{21} & K_{22} & K_{23} & K_{24} \\
K_{31} & K_{32} & K_{33} & K_{34} \\
K_{41} & K_{42} & K_{43} & K_{44} + \gamma
\end{bmatrix}
\begin{bmatrix}
\tilde{\psi}_1 \\
\tilde{\psi}_2 \\
\tilde{\psi}_3 \\
\tilde{\psi}_4
\end{bmatrix}
= \begin{bmatrix}
b_1 \\
b_2 \\
b_3 \\
b_4 + q
\end{bmatrix} + \begin{bmatrix}
g_1 \\
o \\
o \\
o
\end{bmatrix}
\]  
(15.124)

In general for a domain with \( N \) nodes, we have
\[ K_{NN} \rightarrow K_{NN} + \gamma = \frac{\alpha^{(M)}}{l^{(M)}} + \beta^{(M)} \frac{l^{(M)}}{3} + \gamma \]  
(15.125)
\[ b_N \rightarrow b_N + q = f^{(M)} \frac{l^{(M)}}{2} + q \]  
(15.126)
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 \( \tilde{\psi}_i \) \((i = 1, 2, 3, \ldots, N)\) and \( M+1 \) equations. However, the Dirichlet boundary condition given in Eq. (15.68) specifies the value of one of these unknowns, specifically \( \tilde{\psi}_1 \). 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
\[
\begin{bmatrix}
1 & 0 & 0 & 0 \\
K_{21} & K_{22} & K_{23} & K_{24} \\
K_{31} & K_{32} & K_{33} & K_{34} \\
K_{41} & K_{42} & K_{43} & K_{44}
\end{bmatrix}
\begin{bmatrix}
\tilde{\psi}_1 \\
\tilde{\psi}_2 \\
\tilde{\psi}_3 \\
\tilde{\psi}_4
\end{bmatrix}
= \begin{bmatrix}
p \\
b_2 \\
b_3 \\
b_4
\end{bmatrix}
\]  
(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
\[ K_{21} \tilde{\psi}_1 + K_{22} \tilde{\psi}_2 + K_{23} \tilde{\psi}_3 + K_{24} \tilde{\psi}_4 = b_2 \]  
(15.128)
Since \( \tilde{\psi}_1 = p \), however, this equation can be recast as
\[ K_{22} \tilde{\psi}_2 + K_{23} \tilde{\psi}_3 + K_{24} \tilde{\psi}_4 = b_2 - K_{21} p \]  
(15.129)
\(^{22}\)Actually the original first equation \( K_{11} \tilde{\psi}_1 + K_{12} \tilde{\psi}_2 + K_{13} \tilde{\psi}_3 + K_{14} \tilde{\psi}_4 = b_1 + g_1 \) is merely reinterpreted as an equation determining \( g_1 \) after the solution \( \{\tilde{\psi}\} \) has been found.
A similar recasting of the other equations leads finally to

\[
\begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & K_{22} & K_{23} & K_{24} \\
0 & K_{32} & K_{33} & K_{34} \\
0 & K_{42} & K_{43} & K_{44}
\end{bmatrix}
\begin{bmatrix}
\tilde{\varphi}_1 \\
\tilde{\varphi}_2 \\
\tilde{\varphi}_3 \\
\tilde{\varphi}_4
\end{bmatrix}
= 
\begin{bmatrix}
p \\
b_2 - K_{21}p \\
b_3 - K_{31}p \\
b_4 - K_{41}p
\end{bmatrix}
\tag{15.130}
\]

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.\footnote{We assume, of course, that the coefficient matrix is not singular.}

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.

\section*{15.11 Using MATLAB to Solve 1D PDEs via an FEM}

\subsection*{15.11.1 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.\footnote{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.} 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 MATLAB statements\footnote{The \texttt{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 MATLAB, all entered values will be floating, so no adjustment is necessary on those parameters.} (BLOCK 1), including an evaluation of the length of the rod and the coordinates of the nodes, are

\begin{verbatim}
M = input( 'Enter number of segments (M): ' ); M = fix( 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 ];
\end{verbatim}

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\footnote{Remember that indices in MATLAB start at one rather than zero, so all MATLAB indices are coincide with the corresponding subscripts on \([K]\).} (BLOCK 2)
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
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 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 inhomogeneities 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 MATLAB statement (BLOCK 6)

\phi = K\backslash b;

Upon return to the MATLAB command prompt, \(\phi\) 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.2, and can be copied from the directory \$HEAD/matlab.\(^{27}\)

### 15.11.2 An Example: Simple Harmonic Motion

We now illustrate the application of fem1d to a simplification of Eq. (15.67), specifically the equation

\[
m \frac{d^2 \varphi}{dt^2} + k \varphi = 0 \quad ; \quad 0 \leq t \leq T
\]

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.\(^{28}\) 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\(^{29}\)

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

\(^{27}\)At some sites, the file may also be located in the MATLAB directory structure such that MATLAB can find it when it is identified only by its name without a prepended path.

\(^{28}\)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.

\(^{29}\)Remember that we have written the command file fem1d.pro to expect \(l = L/M\).
no further interaction with the program and, presently, the MATLAB 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 MATLAB’s main prompt. In particular, we could plot the solution by invoking the statements

```matlab
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

```matlab
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

```matlab
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', 2);
hold on
plot( x050, phi050, 'Color', 'black', 'LineWidth', 2);
plot( x100, phi100, 'Color', 'black', 'LineWidth', 2);
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:

```matlab
for i=1:50 diff(i) = phi100(2*i) - phi050(i); end
[max(diff), min(diff)]
ans = 0.1444 -0.1490
[max(phi100), min(phi100)]
ans = 1.6035 -1.6045
```

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:

```matlab
(phi100(2)-phi100(1))/ ... Display initial velocity
(x100(2)-x100(1))
-1.3875
(phi100(101)-phi100(100))/ ... Display final velocity
(x100(101)-x100(100))
0.9570
```

---

30This loop stops at $i = 50$ because going to $i = 51$ ultimately seeks $\phi(102)$, which is non-existent.
The results are, of course, in m/s. 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 $0.9789$ m/s, respectively; had we used $M = 500$, we would have found those velocities to be $-1.3857$ and $0.9916$ 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),

$$\frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}$$

which is to be solved subject both to the initial conditions

$$u(x, 0) = f(x) \quad ; \quad \frac{\partial u}{\partial t}(x, 0) = g(x)$$

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

$$u(0, t) = 0 \quad ; \quad \frac{\partial u}{\partial x}(L, t) = 0$$

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

$$\frac{d^2 u_i}{dt^2} = \frac{c^2}{\Delta x^2} \left( u_{i+1} - 2u_i + u_{i-1} \right)$$
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

\[
\frac{d^2 u_0}{dt^2} = 0 ; \quad u_0(0) = 0 , \quad \frac{du_0}{dt}(0) = 0 \tag{15.136}
\]

To address Eq. (15.135) at \( i = N \), we look to a central difference approximation of the spatial derivative, finding that

\[
\frac{\partial u}{\partial x}(L,t) \approx \frac{u_{N+1}(t) - u_{N-1}(t)}{2\Delta x} = 0 \quad \Rightarrow \quad u_{N+1} = u_{N-1} \tag{15.137}
\]

Then, we write Eq. (15.135) for \( i = N \), substitute \( u_{N-1} \) for \( u_{N+1} \), and solve the result for \( \frac{d^2 u_N}{dt^2} \). 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

\[
\frac{d^2 u_0}{dt^2} = 0 \tag{15.138}
\]

\[
\frac{d^2 u_i}{dt^2} = \frac{c^2}{\Delta x^2} \left( u_{i+1} - 2u_i + u_{i-1} \right) , \quad i = 1, 2, 3, \ldots, N - 1 \tag{15.139}
\]

\[
\frac{d^2 u_N}{dt^2} = \frac{c^2}{\Delta x^2} \left( -2u_N + 2u_{N-1} \right) \tag{15.140}
\]

of ODEs subject to the initial conditions

\[
u_i(0) = f(x_i) ; \quad \frac{du_i}{dt}(0) = g(x_i) \tag{15.141}
\]

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 \( t_j = j \Delta t \), \( j = 0, 1, 2, \ldots \). 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

\[
\frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{\Delta t^2} = \frac{c^2}{\Delta x^2} \left( u_{i+1,j} - 2u_{i,j} + u_{i-1,j} \right) \tag{15.142}
\]

Solving for \( u_{i,j+1} \) and introducing \( \alpha = c^2 / \Delta t^2 / \Delta x^2 \), we find the simpler form

\[
u_{i,j+1} = \alpha u_{i+1,j} + 2(1 - \alpha)u_{i,j} + \alpha u_{i-1,j} - u_{i,j-1} \tag{15.143}
\]

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

\[
u_{0,j} = 0 \quad \text{and} \quad \frac{\partial u}{\partial x}(L,t_j) \approx \frac{u_{N+1,j} - u_{N-1,j}}{2 \Delta x} = 0 \quad \Rightarrow \quad u_{N+1,j} = u_{N-1,j} \tag{15.144}
\]
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

\[
  u_{i,0} = f(x_i) \quad \text{and} \quad u_{i,1} = u_{i,0} + g(x_i) \Delta t
\]

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),

\[
\frac{\partial u}{\partial t} = \alpha \frac{\partial^2 u}{\partial x^2}
\]
which is to be solved subject to the initial condition
\[ u(x,0) = f(x) \]  
(15.150)

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
\[ u(0,t) = T_0 \quad \text{and} \quad \frac{\partial u}{\partial x}(L,t) = 0 \]  
(15.151)

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
\[ \frac{du_i}{dt} = \frac{\alpha^2 c}{\Delta x^2} \left( u_{i+1} - 2u_i + u_{i-1} \right) \]  
(15.152)

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
\[ \frac{du_0}{dt} = 0 \quad ; \quad u_0(0) = T_0 \]  
(15.153)

To address Eq. (15.152) at \( i = N \), we look to a central difference approximation of the spatial derivative, finding that
\[ \frac{\partial u}{\partial x}(L,t) \approx \frac{u_{N+1}(t) - u_{N-1}(t)}{2\Delta x} = 0 \quad \Rightarrow \quad u_{N+1} = u_{N-1} \]  
(15.154)

Then, we write Eq. (15.152) for \( i = 0 \), substitute \( u_{N-1} \) for \( u_{N+1} \), and solve the result for \( du_0/dt \).

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
\[ \frac{du_0}{dt} = 0 \]  
(15.155)
\[ \frac{du_i}{dt} = \frac{\alpha^2 c}{\Delta x^2} \left( u_{i+1} - 2u_i + u_{i-1} \right) , \quad i = 1, 2, 3, \ldots, N - 1 \]  
(15.156)
\[ \frac{du_N}{dt} = 2\frac{\alpha^2 c}{\Delta x^2} \left( u_{N-1} - u_N \right) \]  
(15.157)

of ODEs subject to the initial condition
\[ u_i(0) = f(x_i) \]  
(15.158)

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.
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

$$
\frac{u_{i,j+1} - u_{i,j}}{\Delta t} = \frac{\alpha^2}{\Delta x^2} \left( u_{i+1,j} - 2u_{i,j} + u_{i-1,j} \right)
$$

(15.159)

Solving for $u_{i,j+1}$ and introducing $\gamma = \alpha^2 \Delta t/\Delta x^2$, we find the simpler form

$$
u_{i,j+1} = \gamma u_{i-1,j} + (1 - 2\gamma)u_{i,j} + \gamma u_{i+1,j}
$$

(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

$$
u_{0,j} = T_0 \quad \text{and} \quad \frac{\partial u}{\partial x}(L, t_j) \approx \frac{u_{N+1,j} - u_{N-1,j}}{2 \Delta x} = 0 \implies u_{N+1,j} = u_{N-1,j}
$$

(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

$$
u_{0,j+1} = T_0
$$

(15.162)

$$
u_{i,j+1} = \gamma u_{i-1,j} + (1 - 2\gamma)u_{i,j} + \gamma u_{i+1,j}, \quad i = 1, 2, \ldots, N - 1
$$

(15.163)

$$
u_{N,j+1} = 2\gamma u_{N-1,j} + (1 - 2\gamma)u_{N,j}
$$

(15.164)

In their turn, the initial conditions start the process by stipulating that

$$
u_{i,0} = f(x_i)
$$

(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.\(^{31}\)\(^{32}\)

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 \((x, y) = (0, 0), (L, 0), (L, L), \) and \((0, L)\) in the \( xy \) plane. Then, so as to illustrate both possible types of boundary condition, we will suppose that

\[
\begin{align*}
    u(x,0) &= 0 & \quad u(x,L) &= 100 & \quad u(0,y) &= 100 \frac{y}{L} & \quad \frac{\partial u}{\partial x}(L,y) &= 0 \\
\end{align*}
\tag{15.167}
\]

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 \( y_j = j \Delta x \) \((j = 0, 1, 2, \ldots, N)\). 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

\[
\begin{align*}
    \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{\Delta x^2} + \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{\Delta x^2} &= 0 \tag{15.168}
\end{align*}
\]

\(^{31}\)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.

\(^{32}\)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.
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.

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

$$u_{i,j} = \frac{1}{4} \left( u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} \right)$$

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 $u(0, y) = 100 y/L$, or $u_{0,j} = 100 y_j/L$, 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

$$\frac{\partial u}{\partial x}(x, y) \bigg|_{x=L} = 0 \implies \frac{u_{N+1,j} - u_{N-1,j}}{2 \Delta x} = 0 \implies u_{N+1,j} = u_{N-1,j}$$

(15.170)
For these nodes, Eq. (15.169) becomes

\[ u_{N,j} = \frac{1}{4} (u_{N+1,j} + u_{N-1,j} + u_{N,j+1} + u_{N,j-1}) \]

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

\[ u_{N,0} = \frac{1}{4} (2u_{N-1,0} + u_{N,1} + u_{N,-1}) \]  

(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

\[
\begin{align*}
  u_{i,j} &= \frac{1}{4} (u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1}) \quad ; \quad 1 \leq i, j \leq N - 1 \\
  u_{i,0} &= 0 \quad ; \quad 1 \leq i \leq N - 1 \\
  u_{i,N} &= 100 \quad ; \quad 1 \leq i \leq N - 1 \\
  u_{0,j} &= 100 \frac{y_j}{L} \quad ; \quad 1 \leq j \leq N - 1 \\
  u_{N,j} &= \frac{1}{4} (2u_{N-1,j} + u_{N,j+1} + u_{N,j-1}) \quad ; \quad 1 \leq j \leq N - 1 \\
  u_{0,0} &= 0 \\
  u_{0,N} &= 100 \\
  u_{N,0} &= 0 \\
  u_{N,N} &= 100 
\end{align*}
\]  

(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.

15.18 Using MATLAB 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 MATLAB.

15.18.1 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 MATLAB, 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);
c   = input( 'Enter speed of propagation (c): ' );
L   = input( 'Enter length of string (L): ' );
```

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

```
dx = L/N;
x = dx*[ 0 : N ];
alpha = c^2*dt^2/dx^2;
alpha
```

Prudence dictates the wisdom of adding the statement

```
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 \( \text{and} \) 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 \( u1 \) will store the past solution, the vector \( u2 \) will store the current solution, and the vector \( u3 \) 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

\[
\begin{align*}
  u_1 &= \text{zeros}( N+1, 1 ); \quad \text{% For past solution} \\
  u_2 &= \text{zeros}( N+1, 1 ); \quad \text{% For current solution} \\
  u_3 &= \text{zeros}( N+1, 1 ); \quad \text{% For next solution}
\end{align*}
\]

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

\[
f(x) = \begin{cases} 
  0 & 0 \leq x \leq \frac{3}{8}L \\
  1 + \cos \frac{8\pi}{L} \left( x - \frac{L}{2} \right) & \frac{3}{8}L \leq x \leq \frac{5}{8}L \\
  0 & \frac{5}{8}L \leq x \leq L 
\end{cases}
\]

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

\[
g(x) = 0
\]

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

\[
b = 8.0\cdot\pi/L \\
\text{for } i=\text{fix}(3\cdot N/8)+1:\text{fix}(5\cdot N/8) \\
  u1(i) = 1.0 + \cos(b\cdot(x(i)-L/2.0)); \\
\text{end} \\
\text{plot( x, u1, 'linewidth', 3, 'color', 'black' )} \\
\text{set(gca, 'ylim', [-2.0,2.0] );}
\]

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 \texttt{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

\[
u2 = u1 \\
\text{plot( x, u2, 'linewidth', 3.0, 'color', 'black' )} \\
\text{set(gca, 'ylim', [-2.0,2.0] );}
\]

to take the first step and plot the solution at that step.

Now, we are ready to code the algorithm that uses Eqs. (15.145)–(15.147) to advance the solution, step by step. Appropriate coding, or at least a first pass at such coding, might be

\[\text{\textsuperscript{33}}\text{The calculation of the range of } i \text{ to be used is complicated. The MATLAB function } \texttt{fix} \text{ will truncate the quantities } 3\cdot N/8 \text{ and } 5\cdot N/8. \text{ We really want the lower limit to be raised rather than truncated. The upper limit can be truncated. Thus, we add 1 to the lower limit.}\]
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/matlab.\textsuperscript{34}

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 first of these problems by introducing a time delay after each plot is produced by inserting the statement

\[
\text{pause( delay )}
\]

after each \texttt{plot} statement and inserting the statements

\[
\text{delay = input( 'Delay between plots in secs (delay): ' );}
\text{pause on;}
\]

at the beginning so the user can specify the desired delay at execution time and so that MATLAB’s ability to use the command \texttt{pause} is turned on.

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

\[
f = \text{input( 'Plot frequency (f): ' ); } f = \text{fix(f)};
\]

\textsuperscript{34}At some sites, the file may also be located in the MATLAB directory structure such that MATLAB can find it when it is identified only by its name without a prepended path.
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 form\textsuperscript{35}

\begin{verbatim}
if f*fix(j/f) == j
    plot( x, u2, 'linewidth', 3, 'color', 'black' );
    set(gca, 'ylim', [-2.0,2.0] );
    pause( delay );
end
\end{verbatim}

(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

\begin{verbatim}
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
\end{verbatim}

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, etc. 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

\textsuperscript{35}We exploit here the fact that the quantity $f*fix(j/f)$ will be equal to $j$ only when $j$ is a multiple of $f$. 
Figure 15.14: Representative frames in motion of a string. This graph was produced with MATLAB.

```matlab
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
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 = 0.1243 -0.1241
```

Here, at the next to the last MATLAB prompt, we save the final values for later plotting and, in the last MATLAB prompt, we display the range of values assumed by the solution at $t = 20$ s. Then, with the input,

```matlab
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 = 0.0437 -0.0437  

and again with the input  

fdmwave1d  
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 = 0.0190 -0.0190  

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 \( N = 400 \).

### 15.18.2 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 MATLAB, we would begin by requesting input of all necessary parameters and assuring that each is stored with the proper data type. The statements
15.18. USING MATLAB TO SOLVE 2D PDES VIA AN FDM

Figure 15.15: Shape of the string at $t = 20$ s for different segment sizes. This graph was produced with MATLAB.

\begin{verbatim}
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): ' );

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

$$
dx = L/N;$$
$$x = dx*[ 0 : N ];$$
$$gamma = alpha^2*dt/dx^2;$$
$$gamma$$

Prudence dictates the wisdom of adding the statement

$$
if gamma > 0.5 
    disp( 'Error: gamma > 0.5; execution halted' );
    return;
end
$$

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 \( u_1 \) will store the current solution, and the vector \( u_2 \) 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

\[
\begin{align*}
\text{u1} &= \text{zeros}( N+1, 1 ); \quad \% \text{For current solution} \\
\text{u2} &= \text{zeros}( N+1, 1 ); \quad \% \text{For next solution}
\end{align*}
\]

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 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

\[
f(x) = \begin{cases} 
0 & 0 \leq x \leq \frac{3}{8}L \\
1 + \cos \frac{8\pi}{L} \left( x - \frac{L}{2} \right) & \frac{3}{8}L \leq x \leq \frac{5}{8}L \\
0 & \frac{5}{8}L \leq x \leq L
\end{cases}
\]

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

\[
\begin{align*}
\text{b} &= 8.0*\text{pi}/\text{L}; \\
\text{for } \text{i}=\text{fix}(3*\text{N}/8)+1:\text{fix}(5*\text{N}/8) \\
&\quad \text{u1}(\text{i}) = 1.0 + \cos(b*(\text{x}(\text{i})-\text{L}/2.0)); \\
\text{end} \\
\text{plot( x, u1, 'linewidth', 3, 'color', 'black' )} \\
\text{title( 't = 0 s' )} \\
\text{set(gca, 'YLim', [ 0.0, 2.0 ] )}
\end{align*}
\]

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

\[
\begin{align*}
\text{for } \text{j}=1:\text{T} \\
&\quad \text{u2}(1) = 0.0; \\
&\quad \text{for } \text{i}=2:\text{N} \\
&\quad \quad \text{u2}(\text{i}) = \text{gamma}\times\text{u1}(\text{i}-1) + (1.0-2.0\times\text{gamma})\times\text{u1}(\text{i}) + \text{gamma}\times\text{u1}(\text{i}+1); \\
&\quad \text{end} \\
&\quad \text{u2}(\text{N}+1) = 2\times\text{gamma}\times\text{u1}(\text{N}) + (1-2\times\text{gamma})\times\text{u1}(\text{N}+1); \\
&\quad \text{plot( x, u1, 'linewidth', 3, 'color', 'black' )}
\end{align*}
\]

\[\text{36}\] The calculation of the range of \( i \) to be used is complicated. Because of integer arithmetic, the quantities \( 3\times8/8 \) and \( 5\times8/8 \) 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 fmdiffus1d.m, is listed in Appendix 15.D.2, and can be copied from the directory $HEAD/matlab.

Two problems emerge when we run the coding developed to this point with a trial set of values, say \( N = 100,\ dt = 0.1,\ T = 25,\ \alpha = 0.1,\ \text{and}\ L = 10\) (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

\[
pause(\ \text{delay}\ );
\]

after each \textit{plot} statement and inserting the statement

\[
delay = \text{input}(\ '\text{Delay between plots in secs (delay): }\ ' );
\]

\[
pause \text{ on};
\]

at the beginning so the user can specify the desired delay at execution time and so that MATLAB’s ability to use the command \textit{pause} is turned on.

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

\[
f = \text{input}(\ '\text{Plot frequency (f): }\ ' );\ f = \text{fix}(f);
\]

to request the number of solutions whose display should be suppressed after a particular plot has been produced and, in addition, we modify the \textit{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 final \textit{plot} statement in the command file with the conditional statement

\[
\text{if } f*\text{fix}(j/f) == j \\
\quad \text{plot(} x, u1, \text{'linewidth', 3, 'color', 'black'})
\text{title(['}t = ',\text{num2str(j)},', s'])}
\quad \text{set(gca, 'YLim', [ 0.0, 2.0 ] )}
\text{pause( delay );}
\text{end}
\]

\[\text{At some sites, the file may also be located in the MATLAB directory structure such that MATLAB can find it when it is identified only by its name without a prepended path.}\]
Figure 15.16: Temperatures in a rod at various times as it approaches equilibrium. Note that the vertical scale after the graph for $t = 0.0 \text{ s}$ has been expanded the better to show the gradual evolution of the temperature distribution. This graph was produced with MATLAB.

(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

```
fdmdiffus1d
Enter number of segments (N): 100
Enter time step (dt): 0.4
Enter number of time steps (T): 4400
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 end of the rod. Energy diffuses out of the rod altogether at $x = 0$, because the “refrigerator” that maintains that end at $u = 0 \degree \text{C}$ 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

```matlab
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, at the last MATLAB prompt, we save the final values for later plotting. Then, with the input,

```matlab
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

```matlab
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

\(^{38}\)Note that, since \( \gamma = \alpha^2 \frac{dt}{dx^2} \), 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.
Figure 15.17: Temperature distribution at $t = 160$ s for different segment sizes. This graph was produced with MATLAB.

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 MATLAB 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

```matlab
diff = zeros(101,1);
for i=1:101 diff(i)=u1_200(2*i-1)-u1_100(i); end
max( diff )
an = 2.2048e-04
```

Remember that the solution for $N = 200$ has twice as many values at the solution for $N = 100$. 

---

Chapter 15. Solving Partial Differential Equations
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 \( \pm 2 \times 10^{-4} \).

### 15.18.3 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 MATLAB, we would begin by requesting input of the two necessary parameters and assuring that each is stored with the proper data type. The statements

```matlab
N = input( 'Enter number of segments (N): ' ); N = fix(N);
L = 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

```matlab
dx = L/N;
x = dx*[ 0 : N ]; y = 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

```matlab
u = zeros(N+1, N+1);
u(1,:) = 100.0;
u(:,1) = 100.0 - 100.0*x/L;
```

In constructing these statements, we have arranged the values in the array \( u \) so that, in the output generated by the statement `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

```
u = 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
```

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 15.18.1 and 15.18.2. 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 entire frame of the square.
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 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 15.18.1 and 15.18.2, solution of this problem will require an iterative approach.

We have already written Eq. (15.174) to support an iterative approach more suitable than that suggested in the previous paragraph.\(^40\) 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.\(^41\) For the moment, let us simply carry out a user-specified number of iterations. Thus, we would add to the read statements the coding

\[
\text{maxits} = \text{input}( \text{'Maximum number of iterations (maxits): '} );
\]

\[
\text{maxits} = \text{fix}( \text{maxits} );
\]

to obtain the desired number of iterations, we would change the print statement above to read

\[
\text{fprintf}( \text{'
Iteration 0'} )
\]

\[
u
\]

so as to label the output and space it away from what precedes and follows\(^42\) and, finally, we would add the multiple loop

\[
\text{for itcnt}=1: \text{maxits}
\]

\[
\text{for } i = 2:N
\]

\[
\text{for } j = 2:N
\]

\[
u(i,j) = 0.25*(u(i+1,j) + u(i-1,j) + u(i,j+1) + u(i,j-1));
\]

\[
u(i,N+1) = 0.25*(2.0*u(N,j) + u(N+1,j+1) + u(N+1,j-1));
\]

\[
\text{fprintf}( \text{'
Iteration %d', itcnt )}
\]

\[
u
\]

to effect the solution. Here, each pass through the outermost loop will effect one iteration in the algorithm described above. A total of \text{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 \( i \) steps 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 fprintf statement displays the solution after the current iteration is completed. A more fully commented command file

\(^40\)This iterative approach has already been described—though for slightly different boundary conditions—in Section 9.3.1 and implemented in MATLAB in Section 9.9.

\(^41\)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.

\(^42\)Omission of a file ID as the first argument in the fprintf command results in the output being directed to the standard output device, i.e., the screen.
containing these statements—and a few others to be discussed in a moment—is named \texttt{fdmlap2d.m},
is listed in Appendix 15.D.3, and can be copied from the directory \texttt{$\$\text{HEAD/matlab}}.

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

\begin{verbatim}
fdmlap2d
Enter number of segments (N):  5
Enter length of side(L):  10
Maximum number of iterations (maxits):  5
\end{verbatim}

obtaining the output

\begin{verbatim}
Iteration 0
u = 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

Iteration 1
u =  100.0000  100.0000  100.0000  100.0000  100.0000  100.0000
     80.0000  45.0000  36.2500  34.0625  33.5156  41.7578
     60.0000  26.2500  15.6250  12.4219  11.4844  16.1816
     40.0000  16.5625  8.0469  5.1172  4.1504  6.1206
     20.0000  9.1406  4.2969  2.3535  1.6260  2.3431
     0  0  0  0  0  0

Iteration 2
u =  100.0000  100.0000  100.0000  100.0000  100.0000  100.0000
     80.0000  60.6250  52.5781  49.6289  50.7178  54.4043
     60.0000  38.2031  27.8125  23.5107  23.6401  26.9513
     40.0000  23.8477  15.2686  11.3208  10.6769  12.6620
     20.0000  12.0361  7.4146  5.0903  4.5276  5.4293
     0  0  0  0  0  0

Iteration 3
u =  100.0000  100.0000  100.0000  100.0000  100.0000  100.0000
     80.0000  67.6953  61.2842  58.8782  59.2307  61.3531
     60.0000  44.8389  36.2256  32.5162  32.3438  34.6757
     40.0000  28.0359  20.7492  17.2581  16.6979  18.3752
     20.0000  13.8626  9.9255  7.9278  7.5138  8.3507
     0  0  0  0  0  0

Iteration 4
u =  100.0000  100.0000  100.0000  100.0000  100.0000  100.0000
     80.0000  71.5308  66.6586  64.6014  64.5746  65.9562
     60.0000  48.9481  42.2180  39.1053  38.7634  40.4645
     40.0000  30.8900  25.0729  22.2010  21.7133  23.0605
     20.0000  15.2039  12.0512  10.4415  10.1264  10.8283
     0  0  0  0  0  0

Iteration 5
\end{verbatim}
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

\[
f = \text{input('Display frequency (f): ')}; f=\text{fix}(f);
\]

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 \text{fprintf} statement everywhere it appears with the conditional statement

\[
\text{if } f*\text{fix}(/text{itcnt}/f) == \text{itcnt}
\text{fprintf(''
Iteration %d', \text{itcnt} )
\text{u}
\text{end}
\]

(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

\[
\text{fdmlap2d}
\text{Enter number of segments (N): 5}
\text{Enter length of side(L): 10}
\text{Maximum number of iterations (maxits): 80}
\text{Display frequency (f): 20}
\]

obtaining the output

**Iteration 0**

\[
\begin{array}{ccccccc}
100.0000 & 100.0000 & 100.0000 & 100.0000 & 100.0000 & 100.0000 \\
80.0000 & 73.9017 & 70.1803 & 68.4650 & 68.2961 & 69.2642 \\
60.0000 & 51.7524 & 46.5277 & 43.9893 & 43.6158 & 44.8891 \\
0 & 0 & 0 & 0 & 0 & 0 \\
\end{array}
\]

**Iteration 20**

\[
\begin{array}{ccccccc}
100.0000 & 100.0000 & 100.0000 & 100.0000 & 100.0000 & 100.0000 \\
80.0000 & 79.8816 & 79.8018 & 79.7600 & 79.7517 & 79.7702 \\
60.0000 & 59.8314 & 59.7178 & 59.6582 & 59.6464 & 59.6728 \\
\end{array}
\]
15.18. USING MATLAB TO SOLVE 2D PDES VIA AN FDM

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{43} 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 input a larger number of segments, a larger number of iterations, and a larger gap between displayed iterates, and specify the plotting of a labeled mesh surface with the statements\textsuperscript{44}

\begin{verbatim}
mesh(x,y,flipud(u), 'edgecolor', 'black' )
xlabel('x','fontsize',16); ylabel('y','fontsize',16);
zlabel('u','fontsize',16);
set(gca,'fontsize',12)
\end{verbatim}

as an alternative to the printing of a table on the screen with fprintf. A more fully commented command file containing these modifications—and a few others to be discussed in a moment—is named fdmlap2d_plot.m, is listed in Appendix 15.D.4, and can be copied from the directory $HEAD/matlab.

\textsuperscript{43}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{44}The up-down flipping of the array is necessary so that the edge along with u = 100 aligns with the largest index rather than the smallest index.
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

```matlab
pause( delay )
```

after each `mesh` statement and inserting at the beginning the statement

```matlab
delay = input( 'Delay between plots in secs (delay): ' );
pause on
```

so the user has control over that delay at execution time and so that MATLAB is configured to respond to the `pause` statement.

In a first execution of this coding, however, the graphical displays are not yet oriented for the easiest viewing. To repair that problem, we exploit the rotation of the surface in the plot window to find a reasonable orientation, read from the screen the azimuth and elevation associated with that view, and add the statement

```matlab
view( [40 10 ] )
```

immediately after each occurrence of the `mesh` command. (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

```matlab
tdmlap2d_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.

---

45 Those numbers are displayed if the left mouse button is depressed when the cursor is somewhere on the surface.
Convergence of the iterative process to a particular tolerance, however, only provides evidence that the \textit{discretized} equations have been solved to a particular accuracy. That characteristic assures us only that \textit{convergence} error is adequately small. (Probably \textit{roundoff} error is also under control at that point.) By itself, \textit{convergence} of the iterative process offers only marginal assurance that \textit{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. \textit{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 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 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 global labeling is done by assigning a specific and unique integer to each node, as in the left panel of Fig. 15.21. 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. 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 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.

---

46In 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.
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

$$
\tilde{\phi}^{(e)}(x, y) = a^{(e)} + b^{(e)}x + c^{(e)}y
$$

(15.178)
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>

The constants are determined so that this equation gives the correct values at all three nodes of any element, i.e., so that

\[
\tilde{\varphi}_0^{(e)} = a^{(e)} + b^{(e)} x_0^{(e)} + c^{(e)} y_0^{(e)}
\]

\[
\tilde{\varphi}_1^{(e)} = a^{(e)} + b^{(e)} x_1^{(e)} + c^{(e)} y_1^{(e)}
\]

\[
\tilde{\varphi}_2^{(e)} = a^{(e)} + b^{(e)} x_2^{(e)} + c^{(e)} y_2^{(e)}
\]

(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\(^{47}\)

\[
\tilde{\varphi}^{(e)}(x, y) = \sum_{j=0}^{2} N_j^{(e)}(x, y) \tilde{\varphi}_j^{(e)}
\]

(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

\[
N_0^{(e)}(x, y) = \frac{1}{\Delta^{(e)}} \begin{vmatrix}
1 & x & y \\
1 & x_0^{(e)} & y_0^{(e)} \\
1 & x_1^{(e)} & y_1^{(e)} \\
1 & x_2^{(e)} & y_2^{(e)}
\end{vmatrix}
\]

(15.182)

\[
N_1^{(e)}(x, y) = \frac{1}{\Delta^{(e)}} \begin{vmatrix}
1 & x & y \\
1 & x_0^{(e)} & y_0^{(e)} \\
1 & x_1^{(e)} & y_1^{(e)} \\
1 & x_2^{(e)} & y_2^{(e)}
\end{vmatrix}
\]

(15.183)

\[
N_2^{(e)}(x, y) = \frac{1}{\Delta^{(e)}} \begin{vmatrix}
1 & x & y \\
1 & x_0^{(e)} & y_0^{(e)} \\
1 & x_1^{(e)} & y_1^{(e)} \\
1 & x & y
\end{vmatrix}
\]

(15.184)

\(^{47}\)See Exercise 15.15.
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 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

\[
  r = - \frac{\partial}{\partial x} \left( \alpha_x \frac{\partial \bar{\phi}}{\partial x} \right) - \frac{\partial}{\partial y} \left( \alpha_y \frac{\partial \bar{\phi}}{\partial y} \right) + \beta \bar{\phi} - f \tag{15.185}
\]

Evaluated for element \( e \), the weighted residuals then are given by the integral

\[
  R_i^{(e)} = \int \int_{\Omega^{(e)}} N_i^{(e)} r \, dx \, dy \quad ; \quad i = 0, 1, 2 \tag{15.186}
\]

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

\[
  R_i^{(e)} = \int \int_{\Omega^{(e)}} N_i^{(e)} \left[ - \frac{\partial}{\partial x} \left( \alpha_x \frac{\partial \bar{\phi}}{\partial x} \right) - \frac{\partial}{\partial y} \left( \alpha_y \frac{\partial \bar{\phi}}{\partial y} \right) + \beta \bar{\phi} - f \right] \, dx \, dy \tag{15.187}
\]

\[
  = - \int \int_{\Omega^{(e)}} N_i^{(e)} \left[ \frac{\partial}{\partial x} \left( \alpha_x \frac{\partial \bar{\phi}}{\partial x} \right) + \frac{\partial}{\partial y} \left( \alpha_y \frac{\partial \bar{\phi}}{\partial y} \right) \right] \, dx \, dy
  + \int \int_{\Omega^{(e)}} N_i^{(e)} \beta \bar{\phi} \, dx \, dy - \int \int_{\Omega^{(e)}} N_i^{(e)} f \, dx \, dy \tag{15.188}
\]

Then, interpreting \( \psi, V_x, \) and \( V_y \) in Eq. (15.232), the identity developed in Appendix 15.E, as \( N_i^{(e)} \), \( \alpha_x \partial \bar{\phi}/\partial x \), and \( \alpha_y \partial \bar{\phi}/\partial y \), respectively, we can rewrite the first integral in Eq. (15.188) so as to recast the entirety of Eq. (15.188) in the form

\[
  R_i^{(e)} = \int \int_{\Omega^{(e)}} \left( \alpha_x \frac{\partial N_i^{(e)}}{\partial x} \frac{\partial \bar{\phi}^{(e)}}{\partial x} + \alpha_y \frac{\partial N_i^{(e)}}{\partial y} \frac{\partial \bar{\phi}^{(e)}}{\partial y} + \beta N_i^{(e)} \bar{\phi}^{(e)} \right) \, dx \, dy
  - \int \int_{\Omega^{(e)}} N_i^{(e)} f \, dx \, dy - \int_{\Gamma^{(e)}} N_i^{(e)} \left( \alpha_x \frac{\partial \bar{\phi}^{(e)}}{\partial x} \hat{i} + \alpha_y \frac{\partial \bar{\phi}^{(e)}}{\partial y} \hat{j} \right) \cdot \hat{n}^{(e)} \, dl \tag{15.189}
\]

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

\[
  R_i^{(e)} = \sum_{j=0}^{2} \bar{\phi}^{(e)}_j \int \int_{\Omega^{(e)}} \left( \alpha_x \frac{\partial N_i^{(e)}}{\partial x} \frac{\partial \bar{\phi}^{(e)}}{\partial x} + \alpha_y \frac{\partial N_i^{(e)}}{\partial y} \frac{\partial \bar{\phi}^{(e)}}{\partial y} + \beta N_i^{(e)} \bar{\phi}^{(e)} \right) \, dx \, dy
  - \int \int_{\Omega^{(e)}} N_i^{(e)} f \, dx \, dy - \int_{\Gamma^{(e)}} N_i^{(e)} \left( \alpha_x \frac{\partial \bar{\phi}^{(e)}}{\partial x} \hat{i} + \alpha_y \frac{\partial \bar{\phi}^{(e)}}{\partial y} \hat{j} \right) \cdot \hat{n}^{(e)} \, dl \tag{15.190}
\]

or, in the matrix form,

\[
  \{ R^{(e)} \} = [K^{(e)}] \{ \bar{\phi}^{(e)} \} - \{ b^{(e)} \} - \{ g^{(e)} \} \tag{15.191}
\]

\(^{48}\)See Exercise 15.15.
where \([\ldots]\) denotes a matrix and \{\ldots\} denotes a vector. The matrix \([K^{(e)}]\) is a \(3 \times 3\) matrix and \(\{R^{(e)}\}, \{\bar{\varphi}^{(e)}\}, \{p^{(e)}\}, \) and \(\{g^{(e)}\}\) are three-element vectors. The matrix elements \(K^{(e)}_{ij}\) and the vector elements \(b^{(e)}_i\) and \(g^{(e)}_i\) can be obtained from Eq. (15.190), namely

\[
K^{(e)}_{ij} = \int \int_{\Omega^{(e)}} \left( \alpha_x \frac{\partial N_i^{(e)}}{\partial x} \frac{\partial N_j^{(e)}}{\partial x} + \alpha_y \frac{\partial N_i^{(e)}}{\partial y} \frac{\partial N_j^{(e)}}{\partial y} + \beta N_i^{(e)} N_j^{(e)} \right) \, dx \, dy
\]

(15.192)

\[
b^{(e)}_i = \int \int_{\Omega^{(e)}} N_i^{(e)} f \, dx \, dy
\]

(15.193)

\[
g^{(e)}_i = \oint_{\Gamma^{(e)}} N_i^{(e)} \left( \alpha_x \frac{\partial \bar{\varphi}^{(e)}}{\partial x} \hat{i} + \alpha_y \frac{\partial \bar{\varphi}^{(e)}}{\partial y} \hat{j} \right) \cdot \hat{n}^{(e)} \, dl
\]

(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

\[
[K^{(e)}] \{\bar{\varphi}^{(e)}\} = \{b^{(e)}\} + \{g^{(e)}\}
\]

(15.195)

Written out in terms of the specific (local) nodes, this equation becomes

\[
\begin{bmatrix}
K^{(e)}_{00} & K^{(e)}_{01} & K^{(e)}_{02} \\
K^{(e)}_{10} & K^{(e)}_{11} & K^{(e)}_{12} \\
K^{(e)}_{20} & K^{(e)}_{21} & K^{(e)}_{22}
\end{bmatrix}
\begin{bmatrix}
\bar{\varphi}_0 \\
\bar{\varphi}_1 \\
\bar{\varphi}_2
\end{bmatrix}
= \begin{bmatrix}
b^{(e)}_0 \\
b^{(e)}_1 \\
b^{(e)}_2
\end{bmatrix}
+ \begin{bmatrix}
g^{(e)}_0 \\
g^{(e)}_1 \\
g^{(e)}_2
\end{bmatrix}
\]

(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 \(\bar{\varphi}\), 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

\[
[K] \{\bar{\varphi}\} = \{b\} + \{g\}
\]

(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

\[
\bar{\varphi}_0^{(0)} = \bar{\varphi}_1 ; \quad \bar{\varphi}_1^{(0)} = \bar{\varphi}_3 ; \quad \bar{\varphi}_2^{(0)} = \bar{\varphi}_0
\]

(15.198)

and, written with \textit{global} identifications on the unknown \(\bar{\varphi}\)'s, Eq. (15.196) for element 0 becomes

\[
\begin{bmatrix}
K^{(0)}_{00} & K^{(0)}_{01} & K^{(0)}_{02} \\
K^{(0)}_{10} & K^{(0)}_{11} & K^{(0)}_{12} \\
K^{(0)}_{20} & K^{(0)}_{21} & K^{(0)}_{22}
\end{bmatrix}
\begin{bmatrix}
\bar{\varphi}_1 \\
\bar{\varphi}_3 \\
\bar{\varphi}_0
\end{bmatrix}
= \begin{bmatrix}
b^{(0)}_0 \\
b^{(0)}_1 \\
b^{(0)}_2
\end{bmatrix}
+ \begin{bmatrix}
g^{(0)}_0 \\
g^{(0)}_1 \\
g^{(0)}_2
\end{bmatrix}
\]

(15.199)
Similarly, the three nodes of element 1 are (in order) 1, 4, and 3. Thus,

\[ \phi_0^{(1)} = \phi_1 \quad ; \quad \phi_1^{(1)} = \phi_4 \quad ; \quad \phi_2^{(1)} = \phi_3 \]  

and Eq. (15.196) for element 1 becomes

\[
\begin{bmatrix}
K_{00}^{(1)} & K_{01}^{(1)} & K_{02}^{(1)} \\
K_{10}^{(1)} & K_{11}^{(1)} & K_{12}^{(1)} \\
K_{20}^{(1)} & K_{21}^{(1)} & K_{22}^{(1)}
\end{bmatrix}
\begin{bmatrix}
\phi_1 \\
\phi_4 \\
\phi_3
\end{bmatrix}
= 
\begin{bmatrix}
b_0^{(1)} \\
b_1^{(1)} \\
b_2^{(1)}
\end{bmatrix}
+ 
\begin{bmatrix}
g_0^{(1)} \\
g_1^{(1)} \\
g_2^{(1)}
\end{bmatrix}
\]  

We have, of course, found six equations constraining only four unknowns, namely \( \phi_0 \), \( \phi_1 \), \( \phi_3 \), and \( \phi_4 \). Two of the six are redundant.\(^{49}\) 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

\[
\begin{bmatrix}
K_{22}^{(0)} & K_{20}^{(0)} & K_{21}^{(0)} & 0 \\
K_{02}^{(0)} & K_{00}^{(0)} & K_{01}^{(0)} & 0 \\
K_{12}^{(0)} & K_{10}^{(0)} & K_{11}^{(0)} & 0 \\
0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
\phi_0 \\
\phi_1 \\
\phi_3 \\
\phi_4
\end{bmatrix}
= 
\begin{bmatrix}
b_0^{(0)} \\
b_1^{(0)} \\
b_2^{(0)} \\
0
\end{bmatrix}
+ 
\begin{bmatrix}
g_0^{(0)} \\
g_1^{(0)} \\
g_2^{(0)} \\
0
\end{bmatrix}
\]  

Similarly, we augment Eq. (15.201) to obtain the result

\[
\begin{bmatrix}
0 & 0 & 0 & 0 \\
0 & K_{00}^{(1)} & K_{02}^{(1)} & K_{01}^{(1)} \\
0 & K_{20}^{(1)} & K_{22}^{(1)} & K_{21}^{(1)} \\
0 & K_{10}^{(1)} & K_{12}^{(1)} & K_{11}^{(1)}
\end{bmatrix}
\begin{bmatrix}
\phi_0 \\
\phi_1 \\
\phi_3 \\
\phi_4
\end{bmatrix}
= 
\begin{bmatrix}
0 \\
b_0^{(1)} \\
b_1^{(1)} \\
b_2^{(1)}
\end{bmatrix}
+ 
\begin{bmatrix}
g_0^{(1)} \\
g_1^{(1)} \\
g_2^{(1)} \\
0
\end{bmatrix}
\]  

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

\[
\begin{bmatrix}
K_{00}^{(0)} & K_{02}^{(0)} & K_{21}^{(0)} & 0 \\
K_{00}^{(0)} + K_{01}^{(1)} & K_{01}^{(0)} + K_{02}^{(1)} & K_{01}^{(1)} & 0 \\
K_{10}^{(1)} + K_{12}^{(1)} & K_{11}^{(1)} + K_{22}^{(1)} & K_{21}^{(1)} & 0 \\
0 & K_{10}^{(1)} & K_{12}^{(1)} & K_{11}^{(1)}
\end{bmatrix}
\begin{bmatrix}
\phi_0 \\
\phi_1 \\
\phi_3 \\
\phi_4
\end{bmatrix}
= 
\begin{bmatrix}
b_0^{(0)} \\
b_0^{(0)} + b_0^{(1)} \\
b_1^{(0)} + b_2^{(1)} \\
b_1^{(1)}
\end{bmatrix}
+ 
\begin{bmatrix}
g_0^{(0)} \\
g_0^{(0)} + g_1^{(1)} \\
g_1^{(0)} + g_2^{(1)} \\
g_1^{(1)}
\end{bmatrix}
\]  

Furthermore, we can expand these equations to the full 9 × 9 set for all of the nodes, finding that,

\(^{49}\)We assume that the equations are not contradictory.
at this stage in the full assembly,

$$
\begin{bmatrix}
K_{22}^{(0)} & K_{20}^{(0)} & 0 & K_{21}^{(0)} & 0 & 0 & 0 & 0 & 0 \\
K_{02}^{(0)} & K_{00}^{(0)} + K_{02}^{(1)} & 0 & K_{01}^{(0)} + K_{02}^{(1)} & K_{01}^{(0)} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
K_{12}^{(0)} & K_{10}^{(0)} + K_{12}^{(1)} & 0 & K_{11}^{(0)} + K_{12}^{(1)} & K_{11}^{(0)} & 0 & 0 & 0 & 0 \\
0 & K_{10}^{(1)} & 0 & K_{12}^{(1)} & K_{11}^{(1)} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{bmatrix}
\begin{bmatrix}
\tilde{\phi}_0 \\
\tilde{\phi}_1 \\
\tilde{\phi}_2 \\
\tilde{\phi}_3 \\
\tilde{\phi}_4 \\
\tilde{\phi}_5 \\
\tilde{\phi}_6 \\
\tilde{\phi}_7 \\
\tilde{\phi}_8 \\
\end{bmatrix}
$$

$$
= \begin{bmatrix}
b_2^{(0)} \\
b_0^{(0)} + b_0^{(1)} \\
b_1^{(0)} + b_2^{(1)} \\
b_1^{(1)} \\
0 \\
0 \\
0 \\
0 \\
0 \\
\end{bmatrix}
+ \begin{bmatrix}
g_2^{(0)} \\
g_0^{(0)} + g_0^{(1)} \\
g_1^{(0)} + g_2^{(1)} \\
g_1^{(1)} \\
0 \\
0 \\
0 \\
0 \\
0 \\
\end{bmatrix}
$$

(15.205)

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 $b_l^{(e)}$ 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 $g_l^{(e)}$ 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 $\tilde{\phi}$, 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 \textit{a priori} knowledge of these quantities unnecessary or specify these quantities explicitly, we begin by noting that the fully assembled vector \( \{g\} \) is

\[
\{g\} = \begin{cases}
  g_2^{(0)} \\
  g_0^{(0)} + g_0^{(1)} + g_2^{(2)} \\
  g_2^{(2)} + g_0^{(3)} \\
  g_1^{(0)} + g_2^{(1)} + g_2^{(4)} \\
  g_1^{(1)} + g_1^{(2)} + g_2^{(3)} + g_0^{(4)} + g_0^{(5)} + g_2^{(6)} \\
  g_1^{(3)} + g_0^{(6)} + g_0^{(7)} \\
  g_1^{(4)} + g_2^{(5)} \\
  g_1^{(5)} + g_1^{(6)} + g_2^{(7)} \\
  g_1^{(7)}
\end{cases}
\]  

(15.206)

Remember, now, that \( g_i^{(e)} \) is given by the line integral

\[
g_i^{(e)}(e) = \oint_{\Gamma} N_i^{(e)} \left( \alpha_x \frac{\partial \phi_i^{(e)}}{\partial x} \mathbf{i} + \alpha_y \frac{\partial \phi_i^{(e)}}{\partial y} \mathbf{j} \right) \cdot \mathbf{n}^{(e)} \, dl
\]

[see Eq. (15.194)] evaluated counterclockwise around the \textit{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 \textit{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 \((i \rightarrow j + j \rightarrow k + k \rightarrow i)\) 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 \textit{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

\[
(4 \rightarrow 3 + 3 \rightarrow 1 + 1 \rightarrow 4) + \\
(4 \rightarrow 1 + 1 \rightarrow 2 + 2 \rightarrow 4) + \\
(4 \rightarrow 2 + 2 \rightarrow 5 + 5 \rightarrow 4) + \\
(4 \rightarrow 5 + 5 \rightarrow 7 + 7 \rightarrow 4) + \\
(4 \rightarrow 7 + 7 \rightarrow 6 + 6 \rightarrow 4) + \\
(4 \rightarrow 6 + 6 \rightarrow 3 + 3 \rightarrow 4)
\]

Next, note that the middle integral in each line is zero, because the interpolation function along that line is identically zero at all points.\(^{50}\) Then, note that each of the other integrals appears twice, traversed once in each direction, and the two together cancel. \textit{This particular sum boils down to zero!} The same behavior characterizes \textit{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

\[
(2 \rightarrow 4 + 4 \rightarrow 1 + 1 \rightarrow 2) + \\
(2 \rightarrow 5 + 5 \rightarrow 4 + 4 \rightarrow 2)
\]

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,

\(^{50}\)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

$$\{g\} = \begin{cases} 
3 \rightarrow 0 + 0 \rightarrow 1 \\
0 \rightarrow 1 + 1 \rightarrow 2 \\
1 \rightarrow 2 + 2 \rightarrow 5 \\
6 \rightarrow 3 + 3 \rightarrow 0 \\
0 \rightarrow 1 + 1 \rightarrow 2 \\
2 \rightarrow 5 + 5 \rightarrow 8 \\
7 \rightarrow 6 + 6 \rightarrow 3 \\
8 \rightarrow 7 + 7 \rightarrow 6 \\
5 \rightarrow 8 + 8 \rightarrow 7 \\
\end{cases}$$

(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 $[K]\{\tilde{\phi}\} = \{b + g\}$ with the equation $\tilde{\phi}_m = p_m$, 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 $p_m$.\(^{51}\)

Then, to complete the incorporation of this boundary condition, we must reflect the influence of this known value $\tilde{\phi}_m$ 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:

$$K_{n0} \tilde{\phi}_0 + K_{n1} \tilde{\phi}_1 + \cdots + K_{nm} \tilde{\phi}_m + \cdots + K_{n7} \tilde{\phi}_7 + K_{n8} \tilde{\phi}_8 = b_n + g_n$$

(15.209)

Since $\tilde{\phi}_m = p_m$, this expression can be recast as

$$K_{n0} \tilde{\phi}_0 + K_{n1} \tilde{\phi}_1 + \cdots + 0\tilde{\phi}_m + \cdots + K_{n7} \tilde{\phi}_7 + K_{n8} \tilde{\phi}_8 = b_n + g_n - K_{nm} p_m$$

(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

\(^{51}\)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.
15.25. USING MATLAB TO SOLVE 2D PDES VIA AN FEM

• 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 \( 3 \to 0, \ 0 \to 1, \) 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.

15.25 Using MATLAB to Solve 2D PDEs via an FEM

15.25.1 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

\[
\varphi(x, 0) = p_2 \quad ; \quad \varphi(x, L) = p_1
\]  

(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 MATLAB statements are:

```matlab
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., \((M+1)^2\), 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:

```matlab
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 \texttt{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 \textit{global} node number with a simple counter. To accomplish the assignment of values to \( x \) and \( y \), we would use the coding:

```matlab
x = zeros(1:N); % Create array to store x values
y = zeros(1:N); % Create array to store y values
c = 0; % Initialize a counter variable
for i = 0:d % Start row number loop
    for j = 0:d % Start column number loop
c = c+1; % Increment counter
    x(c) = i*dx; % Find x coordinate
    y(c) = 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 MATLAB. 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:

\[ ^{52} \text{The \texttt{fix} function guarantees that parameters that should be integers are, in fact, integers for the subsequent calculations, regardless of how they happen to be entered. In MATLAB, all entered values will be floating, so no adjustment is necessary on those parameters.} \]
Figure 15.22: General element and global node numbering based on the number of segments $d$ into which each edge is divided. Consistent with MATLAB 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 MATLAB.


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 an element number minus 1 [i.e., \((e - 1)\)] by \(2 \times d\) yields 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 MATLAB 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 zeroth 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.\(^{53}\)

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 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 (element number odd)
        cm(2,e) = cm(1,e) + d + 1; % find global number of node 2
        cm(3,e) = cm(1,e) + d; % find global number of node 3
    else % If e is odd (element number even)
        cm(2,e) = cm(1,e) + d; % find global number of node 2
        cm(3,e) = cm(1,e) - 1; % find global number of node 3
    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 \(x_1^{(e)}\) and \(x_2^{(e)}\) in terms of \(x_0^{(e)}\) and \(\Delta x\),

\(^{53}\)Several invocations of the `fix` command are necessary to make sure certain calculated values are integers.
and for $y_1^{(e)}$ and $y_2^{(e)}$ in terms of $y_0^{(e)}$ 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

$$x_2^{(e)} = x_1^{(e)} + \Delta x, \quad x_3^{(e)} = x_1^{(e)} \quad \text{and} \quad y_2^{(e)} = y_3^{(e)} = y_1^{(e)} + \Delta x \quad (e \text{ odd})$$ (15.214)

In contrast, for even numbered elements, we set

$$x_2^{(e)} = x_3^{(e)} = x_1^{(e)} + \Delta x, \quad y_2^{(e)} = y_1^{(e)} \quad \text{and} \quad y_3^{(e)} = y_1^{(e)} + \Delta x \quad (e \text{ even})$$ (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

$$\int \int_{\Omega^{(e)}} \ldots dx dy = \int_{x_1^{(e)}}^{x_1^{(e)} + \Delta x} \left[ \int_{y_1^{(e)}}^{y_1^{(e)} + x - x_1^{(e)}} \ldots dy \right] dx \quad (e \text{ even})$$ (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

$$\int \int_{\Omega^{(e)}} \ldots dx dy = \int_{x_1^{(e)}}^{x_1^{(e)} + \Delta x} \left[ \int_{y_1^{(e)} + x - x_1^{(e)}}^{y_1^{(e)} + \Delta x} \ldots dy \right] dx \quad (e \text{ odd})$$ (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. 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 odd elements and one for even elements, namely

$$K_{\text{even}}^{(e)} = \begin{bmatrix}
\frac{\beta \Delta x^2 + 6\alpha_x}{12} & \frac{\beta \Delta x^2 - 12\alpha_x}{24} & \frac{\beta \Delta x^2}{24} \\
\frac{\beta \Delta x^2 - 12\alpha_x}{24} & \frac{\beta \Delta x^2 + 6\alpha_x + 6\alpha_y}{12} & \frac{\beta \Delta x^2 - 12\alpha_y}{24} \\
\frac{\beta \Delta x^2}{24} & \frac{\beta \Delta x^2 - 12\alpha_y}{24} & \frac{\beta \Delta x^2 + 6\alpha_y}{12}
\end{bmatrix}$$ (15.218)

and

$$K_{\text{odd}}^{(e)} = \begin{bmatrix}
\frac{\beta \Delta x^2 + 6\alpha_y}{12} & \frac{\beta \Delta x^2}{24} & \frac{\beta \Delta x^2 - 12\alpha_y}{24} \\
\frac{\beta \Delta x^2}{24} & \frac{\beta \Delta x^2 + 6\alpha_x}{12} & \frac{\beta \Delta x^2 - 12\alpha_x}{24} \\
\frac{\beta \Delta x^2 - 12\alpha_y}{24} & \frac{\beta \Delta x^2 - 12\alpha_x}{24} & \frac{\beta \Delta x^2 + 6\alpha_x + 6\alpha_y}{12}
\end{bmatrix}$$ (15.219)

The coding that constructs these two matrices in our program is

---

54Details 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.\(^{55}\) 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

\[
\begin{align*}
K &= \text{zeros}(N,N); & \% \text{Create arrays to store values of } K(i,j) \\
b &= \text{zeros}(N,1); & \% \text{K(i,j) and b(i)} \\
\text{for } e = 1:M & \% \text{Count through element numbers} \\
\text{for } i = 1:3 & \% \text{For each local node of an element} \\
\text{b(cm(i,e))} &= \text{b(cm(i,e))} + f*\Delta x^2/6; & \% \text{place its contributions at the} \\
\text{for } j = 1:3 & \% \text{correct locations in } [K] \text{ and } \{b\} \\
\text{if } 2*\text{fix}(e/2) &= e \\
K(cm(i,e),cm(j,e)) &= K(cm(i,e),cm(j,e)) + \text{Kodd}(i,j); & \% \text{K(cm(i,e),cm(j,e))} \\
\text{else} \\
K(cm(i,e),cm(j,e)) &= K(cm(i,e),cm(j,e)) + \text{Keven}(i,j); & \% \text{K(cm(i,e),cm(j,e))} \\
\text{end} \\
\text{end} \\
\text{end}
\end{align*}
\]

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

\(^{55}\)See again Appendix 15.F.
15.25. USING MATLAB TO SOLVE 2D PDES VIA AN FEM

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 \( \tilde{\phi} \), 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
        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
            b(j) = b(j) - K(s,j)*p1; % Reflect influence % of known values of phi in the other
        end
        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 \( g_i^{(e)} \). These two items and the fact that all elements with boundaries on the right edge have even numbers reduce Eq. (15.194) to

\[
g_i^{(e)} = \int_{y_i^{(e)}}^{y_i^{(e)} + \Delta x} N_i^{(e)} \alpha_x \frac{\partial \tilde{\phi}}{\partial x} \, dy
\]  

(15.220)

Since we know that \( \alpha_x(\partial \tilde{\phi}/\partial x) = q \) 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,\(^57\) 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)

\(^{56}\)Remember that indices in MATLAB start at 1. We implement that wrinkle by adding 1 to \( u \), \( s \), and \( t \) and by running the loop on \( j \) from 1 to \( N \) rather than 0 to \( N-1 \).

\(^{57}\)See Appendix 15.F once more.
CHAPTER 15. SOLVING PARTIAL DIFFERENTIAL EQUATIONS

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 \( g^{(c)}_1 \) 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

\[
\begin{align*}
g &= \text{zeros}(N,1); & \text{Create vector to store } g \text{ values} \\
\text{for } i = 1:d-1 & \\
\quad \text{nd} = d*(d+1)+i+1; & \text{Nodes on right boundary} \\
\quad g(\text{nd}) &= q \ast dx; \\
\text{end} & \\
b &= b+g; & \text{Store the vector } \{b+g\} \text{ in } \{b\}
\end{align*}
\]

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

\[
\begin{align*}
\phi &= K\backslash b; & \text{Solve the equation } K\phi = b \\
&\quad \% \text{ writing the results to } \phi \\
A &= \text{zeros}(d+1,d+1); & \text{Create a } d+1 \text{ by } d+1 \text{ array} \\
\text{cnt}=0; & \text{Initialize a counter} \\
\text{for } i = 1:d+1 & \\
\quad \text{for } j = d+1:-1:1 & \text{Use nested for loops to write all} \\
\quad \quad \text{cnt}=\text{cnt}+1; & \text{Increment counter} \\
\quad \quad A(i,j) &= \phi(\text{cnt}); & \text{End inner loop} \\
\text{end} & \text{End outer loop}
\end{align*}
\]

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 MATLAB code has been incorporated into the procedure \texttt{fem2d.m}, a listing of which can be found in Appendix 15.G.2. (The file itself can be copied from the directory \$HEAD/matlab.) 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.

15.25.2 An Example: Isotropic Heat Flow

We will now demonstrate \texttt{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
\[
\begin{align*}
    u(x,0) &= 0 ; \\
    u(x,10) &= 100 ; \\
    u(0,y) &= 10y ; \\
    \frac{\partial u}{\partial x}(10,y) &= 0
\end{align*}
\]  
(15.224)
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

\[
\begin{align*}
    [x, y] &= \text{meshgrid}(0:1.25:10, 0:1.25:10) ; \\
    \text{mesh}(x, y, A', 'edgecolor', 'black', 'linewidth', 3) ; \\
    \text{set(gca,'fontsize',12)} \\
    \text{xlabel('x', 'fontsize',20); ylabel('y', 'fontsize',20)} \\
    \text{zlabel('u','fontsize',20)} \\
    \text{view([45.0,20.0])}
\end{align*}
\]

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\(^{58}\)

\[
\text{rot90( A )}
\]

\[
\begin{align*}
    \text{ans} &= \begin{bmatrix}
        100.00 & 100.00 & 100.00 & 100.00 & 100.00 & 100.00 & 100.00 & 100.00 \\
        87.50 & 87.50 & 87.50 & 87.50 & 87.50 & 87.50 & 87.50 & 87.50 \\
        75.00 & 75.00 & 75.00 & 75.00 & 75.00 & 75.00 & 75.00 & 75.00 \\
        62.50 & 62.50 & 62.50 & 62.50 & 62.50 & 62.50 & 62.50 & 62.50 \\
        50.00 & 50.00 & 50.00 & 50.00 & 50.00 & 50.00 & 50.00 & 50.00 \\
        37.50 & 37.50 & 37.50 & 37.50 & 37.50 & 37.50 & 37.50 & 37.50 \\
        25.00 & 25.00 & 25.00 & 25.00 & 25.00 & 25.00 & 25.00 & 25.00 \\
        12.50 & 12.50 & 12.50 & 12.50 & 12.50 & 12.50 & 12.50 & 12.50 \\
        0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00
    \end{bmatrix}
\end{align*}
\]

reveals even more clearly this anticipated linearity.

\(^{58}\)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 two 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

\begin{verbatim}
 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
\end{verbatim}

which will generate a solution on a 17 × 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 × 9 grid with the statements

\begin{verbatim}
 phi = K\b;
 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{verbatim}
Then, we could extract a 9 × 9 array from this more refined 17 × 17 solution and display that portion of the solution on the original grid with the statements

\[
\begin{align*}
B &= A(1:2:17, 1:2:17); \\
\text{rot90}(B) \\
\end{align*}
\]

\[
\begin{array}{cccccccccc}
100.00 & 100.00 & 100.00 & 100.00 & 100.00 & 100.00 & 100.00 & 100.00 & 100.00 \\
87.50 & 87.50 & 87.50 & 87.50 & 87.50 & 87.50 & 87.50 & 87.50 & 87.50 \\
75.00 & 75.00 & 75.00 & 75.00 & 75.00 & 75.00 & 75.00 & 75.00 & 75.00 \\
62.50 & 62.50 & 62.50 & 62.50 & 62.50 & 62.50 & 62.50 & 62.50 & 62.50 \\
50.00 & 50.00 & 50.00 & 50.00 & 50.00 & 50.00 & 50.00 & 50.00 & 50.00 \\
37.50 & 37.50 & 37.50 & 37.50 & 37.50 & 37.50 & 37.50 & 37.50 & 37.50 \\
25.00 & 25.00 & 25.00 & 25.00 & 25.00 & 25.00 & 25.00 & 25.00 & 25.00 \\
12.50 & 12.50 & 12.50 & 12.50 & 12.50 & 12.50 & 12.50 & 12.50 & 12.50 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
\end{array}
\]

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.)

\section*{15.30 Exercises}

\textit{Note:} In these exercises, we refer to the programs developed in the text \textit{without} appending the file type, adopting this approach to leave to you the choice of which language to use in addressing the exercise.

\subsection*{15.1.} Recast \texttt{fdm1d} to solve the ODE of Eq. (pde:diffeq) when \( f(x) = kx \), the boundary conditions are

\[
\frac{\partial \varphi}{\partial x}(0) = p \quad ; \quad \varphi(L) = q
\]

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

\[
\alpha = -4.0 \quad ; \quad \beta = 4.0 \quad ; \quad L = 10.0 \quad ; \quad p = q = 0
\]

\textbf{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).

\subsection*{15.2.} Suppose the one-dimensional string of length \( l \) discussed in Section 15.1.1 hangs \textit{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

\[
\rho(x) \frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x} \left( \tau(x) \frac{\partial u}{\partial x} \right) \quad ; \quad \tau(x) = -g \int_{x}^{l} \rho(x') \, dx'
\]

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 \( \tau(x) = \rho g (l - x) \). 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 $\bar{x} = x/l$ and $\bar{t} = t/\sqrt{g/l}$.  
(d) Suppose you seek a sinusoidal solution for which $u(\bar{x}, \bar{t}) = f(\bar{x})\cos\omega\bar{t}$. Find the ODE satisfied by $f(\bar{x})$ and then introduce the new independent variable $y$ defined by $y^2 = 1 - \bar{x}$ to find that, expressed as a function of $y$, $f$ satisfies the Bessel equation

$$y^2 \frac{d^2f}{dy^2} + y \frac{df}{dy} - 4\omega^2 y^2 f = 0$$

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 $\alpha(x)$, $\gamma(x)$, 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 \texttt{fdmid} to solve the equation

$$\frac{d^2u}{dx^2} + kux = 0$$

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

$$\nabla^2 u = \frac{1}{\nabla} \frac{\partial}{\partial \nabla} \left( \nabla \frac{\partial u}{\partial \nabla} \right) + \frac{1}{\nabla^2} \frac{\partial^2 u}{\partial \phi^2} = 0 \implies \nabla^2 \frac{\partial^2 u}{\partial \nabla^2} + \nabla \frac{\partial u}{\partial \nabla} + \frac{\partial^2 u}{\partial \phi^2} = 0$$

Suppose a solution is sought in the region $0 < a \leq \nabla \leq b$, $0 \leq \phi \leq 2\pi$ subject to the boundary conditions

$$u(a, \phi) = f(\phi) \quad ; \quad u(b, \phi) = g(\phi)$$

and the requirement that the solution be periodic with period $2\pi$ in $\phi$. Let $\Delta\nabla = (b - a)/n$ and $\Delta\phi = 2\pi/n$ 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 $\partial u/\partial \nabla$. You should find ultimately that

$$u_{i,j} = \frac{u_{i+1,j} \left( \nabla_i^2 \Delta\phi^2 + \frac{1}{2} \nabla_i \Delta\phi^2 \Delta\nabla \right) + u_{i-1,j} \left( \nabla_i^2 \Delta\phi^2 - \frac{1}{2} \nabla_i \Delta\phi^2 \Delta\nabla \right) + (u_{i,j+1} + u_{i,j-1}) \Delta\nabla^2}{2 \nabla_i^2 \Delta\phi^2 + 2 \Delta\nabla^2}$$

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

$$f(\phi) = 0 \quad ; \quad g(\phi) = \begin{cases} -100 & -\pi < \phi < 0 \\ 100 & 0 \leq \phi \leq \pi \end{cases}$$

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

$$\left. \frac{df}{dx} \right|_{x=x_0} = -\frac{3f_0 + 4f_1 - f_2}{2\Delta x}$$
which is a higher-order forward difference approximation than the one used in the text. Hint: Set up the three equations $f_i = Ax_i^2 + Bx_i + C$, 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

$$\frac{d^2u}{dx^2} + k^2 u = 0 \quad ; \quad u(0) = u(L) = 0$$

(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 difference approach to the problem leads to the matrix eigenvalue problem $Au = -k^2L^2u/N^2$ where $N$ is the number of equal-length segments into which the length $L$ of the domain is divided and

$$A = \begin{pmatrix} -2 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -2 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 1 & -2 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 & -2 \end{pmatrix} ; \quad u = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{N-3} \\ u_{N-2} \\ u_{N-1} \end{pmatrix}$$

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

$$\frac{d^2u}{dx^2} + k^2 u = 0 \quad ; \quad u(0) = u(L) = 0$$

(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 $Au = -k^2L^2Bu/6N^2$ where $N$ is the number of equal-length elements into which the length $L$ of the domain is divided and

$$A = \begin{pmatrix} -2 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -2 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 1 & -2 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 & -2 \end{pmatrix} ; \quad B = \begin{pmatrix} 4 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 1 & 4 & 1 & \cdots & 0 & 0 & 0 \\ 0 & 1 & 4 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 4 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 4 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 & 4 \end{pmatrix}$$

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 $x_1(e)$, $x_2(e)$, and $x_3(e)$. 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 $x = x_1(e)$, $x_2(e)$, and $x_3(e)$, respectively. (b) Substituting these values into the above expression and grouping terms appropriately, cast the result in the form

$$\tilde{\varphi}(e) = \sum_{i=1}^{3} \tilde{\varphi}_i(e) N_i(e)(x)$$

and show that the shape functions $N_i(e)(x)$ appropriate to this three-noded linear element are

$$N_1(e)(x) = \frac{1}{\Delta} \begin{vmatrix} 1 & x & x^2 \\ x_2(e) & x_2(e)^2 & x_2(e)^3 \\ x_3(e) & x_3(e)^2 & x_3(e)^3 \end{vmatrix} = \frac{(x_2(e) - x)(x_3(e) - x)}{(x_2(e) - x_1(e))(x_3(e) - x_1(e))}$$

$$N_2(e)(x) = \frac{1}{\Delta} \begin{vmatrix} 1 & x & x^2 \\ x_1(e) & x_1(e)^2 & x_1(e)^3 \\ x_3(e) & x_3(e)^2 & x_3(e)^3 \end{vmatrix} = \frac{(x - x_1(e))(x_2(e) - x)}{(x_2(e) - x_1(e))(x_2(e) - x_3(e))}$$

$$N_3(e)(x) = \frac{1}{\Delta} \begin{vmatrix} 1 & x & x^2 \\ x_1(e) & x_1(e)^2 & x_1(e)^3 \\ x_2(e) & x_2(e)^2 & x_2(e)^3 \end{vmatrix} = \frac{(x - x_1(e))(x_3(e) - x)}{(x_3(e) - x_1(e))(x_3(e) - x_2(e))}$$

where

$$\Delta = \begin{vmatrix} 1 & x_1(e) & x_1(e)^2 \\ 1 & x_2(e) & x_2(e)^2 \\ 1 & x_3(e) & x_3(e)^2 \end{vmatrix}$$

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

$$1^{st}K_{ij} = \int_{x_1(e)}^{x_3(e)} \frac{dN_i}{dx} \frac{dN_j}{dx} dx \quad \text{and} \quad 2^{nd}K_{ij} = \int_{x_1(e)}^{x_3(e)} N_i N_j dx$$

and then construct the $3 \times 3$ matrix whose elements are

$$K_{ij} = a(e) 1^{st}K_{ij} + b(e) 2^{nd}K_{ij}$$
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).

Recast femid 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 \( f(x) = Ae^{-\sigma(x-L/2)^2} \), 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 with a source that provides a constant energy input.

Repeat the previous problem 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.

(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 \( N^{(c)} \) to zero, thereby obtaining the equation of a line in the plane. Verify that that 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 \(\tilde{\varphi} = a + bx + cy + dxy\). (a) Find the four shape functions in general terms, showing that those functions can be expressed in the form

\[
N_0(x, y) = \frac{1}{\Delta} \begin{vmatrix} 1 & x & y & xy \\ 1 & x_1 & y_1 & x_1y_1 \\ 1 & x_2 & y_2 & x_2y_2 \\ 1 & x_3 & y_3 & x_3y_3 \end{vmatrix}
\]

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. 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 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 \(u(0, t) = u(l, t) = 0\) and the initial conditions \(u(x, 0) = 0, \frac{\partial u(x, 0)}{\partial t} = 0\). 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 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 \(u(0, t) = u(l, t) = 0\) 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 that
\[
\frac{u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - \Delta x^2 r_{i,j}}{4 - k^2 \Delta x^2}
\]
Here, \( u_{i,j} = u(x_i, y_j) \), the spacing between consecutive values of \( x \) is \( \Delta x \), the spacing between consecutive values of \( y \) is \( \Delta y \) and \( \Delta y = \Delta x \). This result could be used in an iterative approach to solving the inhomogeneous Helmholtz equation. \textit{Note:} If \( k^2 = 0 \), the equation of this exercise reduces to the inhomogeneous Laplace equation, i.e., to the Poisson equation. If, on the other hand, \( r(x, y) = 0 \), then this equation reduces to the (homogeneous) Helmholtz equation.

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 \textit{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 \textit{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 \( 1 \leq \alpha < 2 \).\(^{59}\) 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 \textit{either} when the tolerance has been reached (or exceeded) \textit{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.

\textit{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 \textit{if} and \textit{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 \textit{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
\[^{59}\alpha \geq 2 \text{ generates an unstable algorithm.}\]
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 fem2d. (c) Find the two-dimensional vector field $V = -\text{grad} U = -\nabla U$ in $R$ and display that field graphically in whatever ways seem appropriate.

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
\[ f(x + h) = f(x) + hf'(x) + \frac{1}{2} h^2 f''(x) + O(h^3) \]
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
\[ u(0, t) = 0 \quad ; \quad u(L, t) = 0 \quad ; \quad u(x, 0) = f(x) \quad ; \quad \left. \frac{\partial u(x, t)}{\partial t} \right|_{t=0} = 0 \]
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 fdmwave1d to solve this problem, testing your program with $f(x) = A \sin \frac{n\pi x}{L}$ for $n = 1, 2, \text{and} 3$. 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.2 Listing of fdm1d.m (MATLAB)

```matlab
% ***** 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 MATLAB’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;

cf = zeros(N+1,N+1);
for i=1:N+1 cf(i,i) = 2.0*alpha + beta*dx2; end
cf(1,1) = 1.0;
for i=1:N cf(i+1,i) = -alpha; end
for i=1:N cf(i,1) = -2.0*alpha;

% ***** Solve system using MATLAB’s operator \.

phi = cf\inhomo;
```
15.B Evaluating Integrals for 1D FEM Problem

15.B.1 ... using MAXIMA

An appropriate batch file to use MAXIMA to evaluate the important integrals in Section 15.1.7 is named FEM1DCalcs.mac, can be copied from the directory $HEAD/maxima, and invoked within MAXIMA with the statement batch( "FEM1DCalcs.mac" ) issued at MAXIMA's prompt for commands. The batch file is listed (without output) in the remainder of this subsection.

/* FEM1DCalcs.mac */
/* 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 */
$n : [ (x2 - x)/l, (x - x1)/l ]$
dnx : diff( n, x )$
/* Create integrand for K, then integrate */
$K1 : transpose(dnx) . dnx$
$K2 : transpose(n) . n$
$K : alpha*K1 + beta*K2$
$K : integrate( K, x, x1, x2 )$
$K : expand( subst( x1+l, x2, K) );$

/* Evaluate b */
$b : integrate( f*n, x, x1, x2 )$
b : expand( subst( x1+l, x2, b ) );$

15.C Program(s) for FEM Approach to 1D Problem

15.C.2 Listing of fem1d.m (MATLAB)

% ***** 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 MATLAB'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): ' ); M = fix( M );
alpa = 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 = 1 * [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
for i = 1:M K(i,i+1) = T; end % main diagonal of K

% 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,1)*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 MATLAB 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);
pause on

% ***** 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*N/8)+1:fix(5*N/8)
    ul(i) = 1.0 + cos(b*(x(i)-L/2.0));
end
plot( x, ul, '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' )
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
pause off;

% ***** 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);
pause on;

% ***** 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;
% ***** Define vectors for current and new solutions

```matlab
u1 = zeros(N+1, 1);  % For current solution
u2 = zeros(N+1, 1);  % For next solution
```

% ***** Set and display initial temperature distribution

```matlab
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] )
```

% ***** Calculate solution repeatedly, displaying every f-th result.

```matlab
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',14)
    set(gca, 'YLim', [0.0, 2.0] )
pause( delay );
  end
  u1 = u2;
end
```

15.D.3 Listing of fdmlap2d.m

% ***** Command file fdmlap2d.m *****

% ***** Acquire controlling parameters *****

```matlab
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 *****

```matlab
dx = L/N;
```
15.D. MATLAB PROGRAMS FOR FDM APPROACH TO 2D PROBLEMS

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 *****

fprintf( '
Iteration 0' )
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
        fprintf( '
Iteration %d', 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): ' );
pause on
%
****** 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,flipud(u), 'edgecolor', 'black' )
xlabel('x','fontsize',16); ylabel('y','fontsize',16);
zlabel('u','fontsize',16);
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,flipud(u), 'edgecolor', 'black' )
xlabel('x','fontsize',16); ylabel('y','fontsize',16);
zlabel('u','fontsize',12);
set(gca,'fontsize',12)
view( [40 10] )
pause( delay )
    end
end
pause off

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

\[ I = \int \int_{\Omega} \psi \left( \frac{\partial V_x}{\partial x} + \frac{\partial V_y}{\partial y} \right) \, dx \, dy \]  

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

\[ I = \int \int_{\Omega} \left( \frac{\partial (\psi V_x)}{\partial x} + \frac{\partial (\psi V_y)}{\partial y} \right) \, dx \, dy - \int \int_{\Omega} \left( V_x \frac{\partial \psi}{\partial x} + V_y \frac{\partial \psi}{\partial y} \right) \, dx \, dy \]  

(15.226)

Now, introduce the vector \( Q \) whose \( x \) and \( y \) components are \( -\psi V_y \) and \( \psi V_x \), respectively. Since

\[ \frac{\partial (\psi V_x)}{\partial x} + \frac{\partial (\psi V_y)}{\partial y} = \frac{\partial Q_y}{\partial x} - \frac{\partial Q_x}{\partial y} = (\nabla \times Q)_z = (\nabla \times Q) \cdot \hat{k} \]  

(15.227)

the integral of concern then can be written in the form

\[ I = \int \int_{\Omega} (\nabla \times Q) \cdot \hat{k} \, dx \, dy - \int \int_{\Omega} \left( V_x \frac{\partial \psi}{\partial x} + V_y \frac{\partial \psi}{\partial y} \right) \, dx \, dy \]  

(15.228)
15.F. Evaluating Integrals for 2D FEM Problem

which, upon invoking Stokes’ theorem, we can rewrite further in the form

\[ I = \int \oint \Gamma Q \cdot \hat{t} \, dl - \int \int \Omega \left( V_x \frac{\partial \psi}{\partial x} + V_y \frac{\partial \psi}{\partial y} \right) \, dx \, dy \]  

(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 \( Q \) 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

\[ \hat{t} = \hat{k} \times \hat{n} = \hat{k} \times (n_x \hat{i} + n_y \hat{j}) = -n_y \hat{i} + n_x \hat{j} \]  

(15.230)

Thus,

\[ Q \cdot \hat{t} = n_x Q_y - n_y Q_x = n_x \psi V_x + n_y \psi V_y = (\psi V_x \hat{i} + \psi V_y \hat{j}) \cdot \hat{n} = \psi (V_x \hat{i} + V_y \hat{j}) \cdot \hat{n} \]  

(15.231)

In other words, the tangential component of \( Q \) can be recast as the normal component of the vector from which \( Q \) was originally derived. Thus, we find that the original integral in Eq. (15.225) can alternatively be evaluated as

\[ I = -\int \int \Omega \left( V_x \frac{\partial \psi}{\partial x} + V_y \frac{\partial \psi}{\partial y} \right) \, dx \, dy + \oint \Gamma (\psi V_x \hat{i} + \psi V_y \hat{j}) \cdot \hat{n} \, dl \]  

(15.232)

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.1 ... using MAXIMA

An appropriate batch file to use MAXIMA to evaluate the necessary integrals when array indices start at 0 is named \texttt{FEM2DCalcs0.mac}, can be copied from the directory \$HEAD/maxima and, with the directory containing this file in MAXIMA’s search path,\(^{60}\) invoked within MAXIMA with the statement \texttt{batch( "FEM2DCalcs0.mac" )} issued at MAXIMA’s prompt for commands. This batch file includes the statements

```maxima
/* FEM2DCalcs0.mac */
/* This batch file evaluates several integrals that appear in */
/* setting up coding to solve the 2D FEM problem in CPSUP */
/* for programs whose array indices start at 0. */
/* Create interpolation functions and their derivatives */

r : [1, x, y]; /* Create rows */
r0 : [1,x0,y0];
r1 : [1,x1,y1];
```

\(^{60}\) Alternatively, the full path to the file can be included within the quotation marks in the indicated command.
r2 : [1, x2, y2]

delta : determinant( matrix( r0, r1, r2 ) )$ /* Evaluate denominator */

n0 : determinant( matrix( r, r1, r2 ) )/delta$ /* Evaluate functions */
n1 : determinant( matrix( r0, r, r2 ) )/delta$
n2 : determinant( matrix( r0, r1, r ) )/delta$
n : [ n0, n1, n2 ]$

/* Simplify functions and evaluate derivatives */
/* for odd-numbered elements */

(x1 : x0 + dx, x2 : x0 + dx, y1 : y0, y2 : y0 + dx )$
nodd : expand( ev(n) )$
dnoddx : diff( nodd, x )$
dnoddy : diff( nodd, y )$

/* Simplify functions and evaluate derivatives */
/* for even-numbered elements */

(x1 : x0 + dx, x2 : x0, y1 : y0 + dx, y2 : y0 + dx )$
neven : expand( ev(n) )$
dnevenx : diff( neven, x )$
dneveny : diff( neven, y )$

remvalue( x1, x2, y1, y2 )$

/* Create integrand for Kodd, then integrate */

Kodd : ax*transpose(dnoddx).dnoddx + ay*transpose(dnoddy).dnoddy +
      b*transpose(nodd).nodd$
Kodd : integrate( integrate(Kodd, y, y0, y0+x-x0 ), x, x0, x0+dx )$
Kodd : expand( Kodd );

/* Create integrand for Keven, then integrate */

Keven : ax*transpose(dnevenx).dnevenx + ay*transpose(dneveny).dneveny +
      b*transpose(neven).neven$
Keven : integrate( integrate(Keven, y, y0+x-x0, y0+dx ), x, x0, x0+dx )$
Keven : expand( Keven );

/* Evaluate bodd and beven */

bodd : integrate( integrate( nodd*f, y, y0, y0+x-x0 ), x, x0, x0+dx )$
bodd : expand( bodd );

beven : integrate( integrate( neven*f, y, y0+x-x0, y0+dx ), x, x0, x0+dx )$
beven : expand( beven );

/* Evaluate needed g’s */

g : integrate( nodd*q, y, y0, y0+dx )$
g : expand( subst( x0+dx, x, g ) );
An appropriate batch file to use MAXIMA to evaluate the necessary integrals when array indices start at 1 is named `FEM2DCalcs1.mac`, can be copied from the directory `$HEAD/maxima` and, with the directory containing this file in MAXIMA’s search path,\(^{61}\) invoked within MAXIMA with the statement `batch("FEM2DCalcs1.mac")` issued at MAXIMA’s prompt for commands. This batch file includes the statements

```plaintext
/* FEM2DCalcs1.mac */

/* This batch file evaluates several integrals that appear in */
/* setting up coding to solve the 2D FEM problem in CPSUP */
/* for programs whose array indices start at 1. */

/* Create interpolation functions and their derivatives */

r : [1, x, y]$

r1 : [1,x1,y1]$

r2 : [1,x2,y2]$

r3 : [1,x3,y3]$

delta : determinant( matrix( r1,r2,r3 ) )$ /* Evaluate denominator */

n1 : determinant( matrix( r,r2,r3 ) )/delta$ /* Evaluate functions */

n2 : determinant( matrix( r1,r,r3 ) )/delta$

n3 : determinant( matrix( r1,r2,r ) )/delta$

n : [ n1, n2, n3 ]$

/* Simplify functions and evaluate derivatives */
/* for even-numbered elements */

(x2 : x1 + dx, x3 : x1 + dx, y2 : y1, y3 : y1 + dx )$

neven : expand( ev(n) )$

dnevenx : diff( neven, x )$

dneveny : diff( neven, y )$

/* Simplify functions and evaluate derivatives */
/* for odd-numbered elements */

(x2 : x1 + dx, x3 : x1, y2 : y1 + dx, y3 : y1 + dx )$

nodd : expand( ev(n) )$

dnoddx : diff( nodd, x )$

dnoddy : diff( nodd, y )$

remvalue( x1, x2, y1, y2 )$

/* Create integrand for Keven, then integrate */

Keven : ax*transpose(dnevenx).dnevenx + ay*transpose(dneveny).dneveny + b*transpose(neven).neven$

Keven : integrate( integrate(Keven, y, y1, y1+x-x1 ), x, x1, x1+dx )$

Keven : expand( Keven );

/* Create integrand for Kodd, then integrate */

```

\(^{61}\)Alternatively, the full path to the file can be included within the quotation marks in the indicated command.
\[ K_{\text{odd}} : ax^{\text{transpose}}(dn_{\text{odd}}x) + ay^{\text{transpose}}(dn_{\text{odd}}y) + \]
\[ b^{\text{transpose}}(n_{\text{odd}}) \]
\[ K_{\text{odd}} : \int \int (K_{\text{odd}}, y, y_1+x-x_1, y_1+dx), x, x_1, x_1+dx) \]
\[ K_{\text{odd}} : \text{expand}(K_{\text{odd}}); \]

/* Evaluate b_{\text{odd}} and b_{\text{even}} */

\[ b_{\text{even}} : \int \int (n_{\text{even}}f, y, y_1, y_1+x-x_1), x, x_1, x_1+dx) \]
\[ b_{\text{even}} : \text{expand}(b_{\text{even}}); \]

\[ b_{\text{odd}} : \int \int (n_{\text{odd}}f, y, y_1+x-x_1, y_1+dx), x, x_1, x_1+dx) \]
\[ b_{\text{odd}} : \text{expand}(b_{\text{odd}}); \]

/* Evaluate needed g's */

\[ g : \int (n_{\text{even}}q, y, y_1, y_1+dx) \]
\[ g : \text{expand}(\text{subst}(x_1+dx, x, g)); \]
15.G  Program(s) for FEM Approach to 2D Problems

15.G.2  Listing of fem2d.m (MATLAB)

% ***** 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 2
        cm(3,e) = cm(1,e) + d; % find global number of node 3
    else % If e is odd
        cm(2,e) = cm(1,e) + d ; % find global number of node 2
        cm(3,e) = cm(1,e) -1; % find global number of node 3
    end
end
% ***** Find [K] for odd and even numbered elements *****

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) = Kodd(2,3);
Kodd(3,1) = Kodd(1,3);

% ***** Create stiffness matrix and vector {b} *****

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

% ***** 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
        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;
        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 $\texttt{\$HEAD/tex}$, where (as defined in the Local Guide) $\texttt{\$HEAD}$ 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 $\TeX^1$ 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, $\LaTeX^2$ 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,$^3$ 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 wysiwyg,$^4$ 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.$^5$ 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.$^6$

---

1As 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’.

2Here, individuals disagree on the pronunciation of the ‘la’, and Leslie Lamport offers no guidance. Some say lah$\TeX$ 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.

3The 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 $\TeX$/$\LaTeX$ system can be downloaded.

4what you see is what you get.

5See Section A.18 for more detailed references.

6You may also find the web pages of TUG at the 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 $\TeX$. 

457
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 \texttt{begin} with a backslash (\texttt{	extbackslash}). \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 any \LaTeX document specifies the document \texttt{class} and has the form\footnote{Be aware that some of the features described in this Appendix are specific to \LaTeX2ε and will not work with the previous version (\LaTeX 2.09). \LaTeX2ε, 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} \quad \texttt{or} \quad \documentclass[options]{Class}
\end{verbatim}

where

- \texttt{Class} is any one of the options shown in Table A.1 and
- \texttt{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}
Table A.1: Document classes valid in standard \LaTeX. Other available classes may be described in the Local Guide.

- **article** This class is by far the most common. It is used for most short documents. See Sections 2.2.2 and C.5.1 in The \LaTeX Manual.
- **report** This class is generally used for longer documents. See Sections 2.2.2 and C.5.1 in The \LaTeX Manual.
- **book** This class is meant for actual books. See Sections 5.1 and C.5.1 in The \LaTeX Manual.
- **slides** This class can be used in creating originals from which overhead transparencies to be used in presentations can be made. The type size, including the size used for mathematical symbols, is quite large, and transparencies readable from some distance will be produced. See Sections 5.2 and C.5.1 in The \LaTeX Manual.
- **letter** This class facilitates creating letters. See Sections 5.3 and C.5.1 in The \LaTeX Manual.

Table A.2: The most common options for the \texttt{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.

- **10pt** or **11pt** or **12pt** (default = 10pt) Specifies default type size.
- **letterpaper** or **legalpaper** or **a4paper** or ... (default depends on site but is usually letterpaper in the US) Specifies size of paper to be used.
- **landscape** (default is portrait orientation) Specifies that text is to be formatted for landscape orientation on the specified paper size.
- **oneside** or **twoside** (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.
- **onecolumn** or **twocolumn** (default = onecolumn) Specifies one- or two-column formatting on each page.

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.

Whether 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.

\begin{verbatim}
\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}
\end{verbatim}

\section*{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

   \verbatim
   \texttt{latex filename}

   (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). 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 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 \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. The resulting .dvi file can be directly displayed on the screen (see Section A.1.4).

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

\footnote{See the \textit{Local Guide} for the precise command at your site.}

\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 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 third—pass is necessary to incorporate that information in the final formatting of the pages.}

\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.}
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

```
dvips -o filename.ps -t letter filename           (Default extension for the input file is .dvi.)
```

which translates the entire file into PostScript, stores it in a file, and (with the -o option here used) gives the stored file the same name as the input file but with extension .ps and with the -t option forces use of letter size paper or (2) by selecting an item from a menu.15

To be more explicit, the file texsample1.tex, contains the L\LaTeX code in Table A.3. Once this file has been copied to the default directory, we would produce the .dvi and .ps files with statements like

```
latex texsample1

dvips -o texsample1.ps -t letter texsample1
```

or equivalent selections from a menu.

A.1.3 Creating a PDF File

PDF files can be created from the L\LaTeX code in at least two ways:

1. Starting with L\LaTeX source file created in Section A.1.1, one submits to the operating system the single statement

   ```
pdflatex filename                          (Default file type is .tex)
```

   or, for the sample above,

   ```
pdflatex texsample1
```

\[\text{12}\text{Additional information about dvips can be found by typing the command dvips with no arguments. Further, the option \texttt{-pp} allows selection of specific pages from the full document, the options \texttt{-p} (first page) and \texttt{-l} (last page) allow selection of a range of pages from the full document, and the option \texttt{-t} allows printing of the output in landscape orientation rather than in the default portrait orientation. Even more detailed information is available in the on-line help for dvips, accessed in many systems by typing the statement texdoc dvips at a Shell prompt or, in the UNIX and LINUX operating systems, the statement man dvips at a Shell prompt.}\]

\[\text{13}\text{If the explicit specification of the output file is omitted, dvips will attempt to send the file directly to the printer—and the whole operation will probably fail. Further, the PostScript file will then not be stored for subsequent use.}\]

\[\text{14}\text{This option is often the default. The option \texttt{-t landscape} will override the default portrait orientation.}\]

\[\text{15}\text{Check the Local Guide for more details. In particular, some sites may have implemented the shorthand command dvip filename for the command dvips -o filename.ps filename.}\]

\[\text{16}\text{The program pdflatex is normally installed automatically whenever L\LaTeX is installed.}\]
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

\begin{verbatim}
ps2pdf filename.ps  \\
\end{verbatim}

(No default file type)

or, for the sample above

\begin{verbatim}
ps2pdf texsample1.ps
\end{verbatim}

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.

### 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.}

### 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 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{verbatim}
UNIX/LINUX/MAC  \\
lp filename.ps
\end{verbatim}

Windows

\begin{verbatim}
print filename.ps
\end{verbatim}

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

\setlength{\textwidth}{Value}


\setlength{\oddsidemargin}{0.25truein}
\setlength{\topmargin}{-0.5truein}
\setlength{\parindent}{20pt}
\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\(^{21}\)) 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\(^{22}\)

\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

---

\(^{20}\)The *point*, which is a standard printer's measure of length, is 1/72” (72 points per inch).

\(^{21}\)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.

\(^{22}\)By convention in LATEX, *optional* arguments to a command are enclosed in *square brackets* and *mandatory* arguments are enclosed in *curly braces*. 

---
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 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 The \LaTeX\ Manual.

\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.

\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.

\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.

\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.

\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 \headheight) and the separation of the header from the main text (specified by \headsep) together add to 0.5", so—if the defaults for \headheight and \headsep are accepted—we can pretend that \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.

\parindent (default 15.0 points) Specifies the paragraph indent.
\parskip (default 0.0 points) Specifies the extra space between paragraphs.

Table A.6: Templates in the directory $\HEAD/tex$.

<table>
<thead>
<tr>
<th>Template</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>lu_article.template</td>
<td>Specifies a general article style, including a 6&quot;×9&quot; text area centered on an 8.5&quot;×11&quot; page.</td>
</tr>
<tr>
<td>lu_cpl.template</td>
<td>Specifies the style of CPL publications.</td>
</tr>
<tr>
<td>lu_sloan.template</td>
<td>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.</td>
</tr>
<tr>
<td>lu_letter.template</td>
<td>Supplies a starting format for business letters.</td>
</tr>
<tr>
<td>lu_memo.template</td>
<td>Supplies a starting format for memos.</td>
</tr>
</tbody>
</table>
Table A.7: Commands for changing type style. *Note:* Some of these commands may exhibit unexpected behaviors when used in math mode and one of them (\textsc{sc}) is invalid in math mode.

<table>
<thead>
<tr>
<th>Language</th>
<th>Command</th>
<th>Function</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>\TeX{}</td>
<td>\text{\textit{text}}</td>
<td>Sets text in medium, Roman, italic type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\TeX{}</td>
<td>\text{\textsf{text}}</td>
<td>Sets text in medium, sans serif, upright type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\LaTeX{}</td>
<td>\text{\texttt{text}}</td>
<td>Sets text in medium, typewriter, upright type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\LaTeX{}</td>
<td>\text{\textbf{text}}</td>
<td>Sets text in bold, Roman, upright type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\LaTeX{}</td>
<td>\text{\textsc{text}}</td>
<td>Sets text in medium, Roman, small caps type.</td>
<td>SAMPLE</td>
</tr>
<tr>
<td>\LaTeX{}</td>
<td>\text{\textit{text}}</td>
<td>Sets text in emphasis type.</td>
<td>(See text.)</td>
</tr>
</tbody>
</table>

*Local Guide* for information about possible additional templates available at your site.

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 caps) 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 \textbf{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 \textit{might be used. The declaration \textup{, however, is preferred because it is sensitive to the current type style. When the current type style is Roman, \textup{ will shift to italic type; when the current type style is italic, \textup{ will shift to Roman type. Thus, \textup{ 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.

---

23The default is medium series, Roman family, upright shape—the type style used for this book.

24Because 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 \emph{declaration} \texttt{\em}, but also with the \texttt{command} \texttt{\emph}. 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}}. (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 \texttt{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 } } }}

\texttt{to achieve the same end with a shorter phrase. The full set of available declarations and commands includes}

\texttt{\mdseries, \bfseries, \textmd{...}, \textbf{...}}

\texttt{\to specify medium and bold series, respectively,}

\texttt{\rmfamily, \sffamily, \ttfamily, \textsf{...}, \texttt{...}}

\texttt{\to specify Roman, sans serif, and typewriter families, respectively, and}

\texttt{\upshape, \itshape, \slshape, \scshape ,
\textup{...}, \textit{...}, \textsl{...}, \textsc{...}}

\texttt{to specify upright, italic, slanted, and small-cap shapes. A quick return to the default specified in the originally invoked document class is accomplished with the declaration \texttt{\normalfont} or the command \texttt{\textnormal}.}

\textbf{A.3.2 Type Size}

Type size is independent of type style. The global type size for a particular document is specified in the command \texttt{\documentclass} as described in Section A.2. A local change in type size is accomplished by one or another of the declarations in Table A.8. The selected type size is determined relative to the global type size specified in the command \texttt{\documentclass}. Again curly braces are used to limit the scope of the change in size.

To combine a change in type \textit{style} with a change in type \textit{size}, place the specification of type size \texttt{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{text} in tiny type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\scriptsize \text{text}</td>
<td>Sets \text{text} in scriptsize type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\footnotesize \text{text}</td>
<td>Sets \text{text} in footnotesize type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\small \text{text}</td>
<td>Sets \text{text} in small type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\normalsize \text{text}</td>
<td>Sets \text{text} in normalsize type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\large \text{text}</td>
<td>Sets \text{text} in large type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\Large \text{text}</td>
<td>Sets \text{text} in Large type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\LARGE \text{text}</td>
<td>Sets \text{text} in LARGE type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\huge \text{text}</td>
<td>Sets \text{text} in huge type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\Huge \text{text}</td>
<td>Sets \text{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 \emph{x}, and an em space is about the width of the letter \emph{M}, both in the current font.)

- The command \vspace{\emph{value}} to generate vertical space. Here, \emph{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*{\emph{value}} will force the space to be included even if it is at the top of a page.

- The command \hspace{\emph{value}} to generate horizontal space. Here, \emph{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*{\emph{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

\begin{verbatim}
\newlength{dblast}
\settowidth{dblast}{**}
\end{verbatim}

and then insert the space for two asterisks (without actually displaying them) with the command

\begin{verbatim}
\hspace{dblast}
\end{verbatim}

These commands are described in detail at the end of Sections 6.4.1 and C.13.1 in *The \LaTeX\ Manual*.

### 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 *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{verbatim}
\begin{quotation}
Text of quotation.
\end{quotation}
\end{verbatim}

or

\begin{verbatim}
\begin{quote}
Text of quotation.
\end{quote}
\end{verbatim}

Normally, the command \begin{verbatim}\begin{...}\end{verbatim} will be placed on a line by itself, the text may spread over several lines, and the command \begin{verbatim}\end{...}\end{verbatim} 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 no extra space between paragraphs while the second will place extra space between paragraphs and will 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{verbatim}
\begin{center}
First line \ \ Second line \ \ Third line \ \ ...
\end{center}
\end{verbatim}

Normally, the command \begin{verbatim}\begin{...}\end{verbatim} 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 \texttt{\textbackslash{}}, and the command \begin{verbatim}\end{...}\end{verbatim} will be placed on a line by itself.

These two examples introduce the construction called an *environment*, which in general has the format

\begin{verbatim}
\begin{EnvironmentName}
...
\end{EnvironmentName}
\end{verbatim}

We met the \texttt{document} environment early on. Beyond the \texttt{quotation}, \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 *The
A.3. 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 \textit{The \LaTeX Manual}—the parameter \texttt{\baselinestretch} is \textit{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}!

A.3.5 \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 \texttt{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 \textit{The \LaTeX Companion} describes numerous packages, including \texttt{amstex}, \texttt{babel}, \texttt{color}, \texttt{graphics}, \texttt{graphicsx}, \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 \textit{The \LaTeX Manual}. The statement \texttt{texdoc PackageName} at a \texttt{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.

A.3.6 Miscellaneous Other Capabilities

Several additional capabilities merit particular comment:

- 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{\~} instead of a space between the words. The tilde should also be used after a period that does \textit{not} end a sentence. For example, we should type ‘Mr.\texttt{\~}Cook’ rather than\texttt{\~} ‘Mr.\texttt{\~}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.

- 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
dash 25The symbol '\_\_' underscores the presence of a space at the indicated point; it is \textit{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 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 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 \TeX{} is permitted to insert a hyphen as it fills and justifies lines. \TeX{} has its own rules for hyphenation, but they are not infallible. Sometimes \TeX{} 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 \mbox command, e.g., \mbox{customize}.

A.4 Including Equations

Most scientific documents will have equations, all of which must be specified in \TeX{}’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 \(...\) or \$...\$, which are equivalent. If we want one or more displayed equations, with or without automatically generated equation numbers, we need to use the form \begin...\end. Three standard environments use this construction. The displaymath environment centers a single equation on a line by itself, does not number the equation, and can be specified by the shortened form \[...\]. The equation environment centers a single equation on a line by itself and numbers the equation automatically. The eqnarray environment—see \emph{The \TeX{} 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 equation and eqnarray environments, \TeX{} provides the command \label{ReferenceName} for defining a symbolic label \textit{within} the environment and the command \ref{ReferenceName} to permit referencing the equation by its number in the text. Further, the command \pageref{ReferenceName} permits referencing the \textit{page} on which the equation occurs in the text.\footnote{When these references are used, the information written into the .aux file becomes important and the code must be processed by \textit{latex} or \textit{pdflatex} \textit{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 \textit{table} and \textit{figure} also admit a \texttt{\label} command and can make use of this capability for symbolic internal references to those components. The \texttt{\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.} (Warning: The displaymath, equation, and eqnarray environments will generate error messages if they contain blank lines.)
LTtEX 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.\footnote{For a complete list of all symbols and the commands that create them, see Section 3.3 on Mathematical Formulas in \textit{The \LaTeX\ Manual}.} To reiterate, these symbols and structures can only be used if LTtEX is in math mode. So, to insert the symbol $\Theta$ in the final output, we place the command $\Theta$ in the code.\footnote{In math mode, Greek letters can be produced by commands that simply name the letter, e.g., \texttt{Theta} for $\Theta$ and \texttt{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., \texttt{Kappa}, do not exist.} (Here, the first dollar sign puts LTtEX in undisplayed math mode and the second returns LTtEX to text mode.) Since LTtEX 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 \texttt{omega} from the character $t$.) The commands $\,, \,!\, ;$ 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 $(x + y)dx\, dy$.$\footnote{Technically and officially, the differential element $dx$ should be written with a \textit{Roman} \texttt{d} and an \textit{italic} \texttt{x}, i.e., $dx$, but this convention is rarely followed.}$

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 \texttt{cos} and \texttt{sin} should be used for these functions.\footnote{See Table 3.9 in \textit{The \LaTeX\ Manual} for a complete listing.} Except for \texttt{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 \texttt{mathrm}, \texttt{mathit}, \texttt{mathbf}, \texttt{mathtt}, etc. described in Section 3.3.8 of \textit{The \LaTeX\ Manual}. Remember, however, that spaces are ignored in math mode, so a shift to Roman type with the command \texttt{mathrm}, for example, to include a two-word phrase, will result in the words running together unless the space command \texttt{\␣} is used to insert an \textit{explicit} space between the words. Sometimes, the better means to insert Roman text in equations is to use the command \texttt{mbox} to escape temporarily to text mode.

Including bold characters in \textit{equations} is especially tricky. The command \texttt{mathbf} “emboldens” only \textit{some} characters in the processed formula and, in particular, leaves lower case Greek symbols in light face. The \texttt{boldmath} declaration causes \textit{everything} in math mode to be bold but cannot itself be \textit{invoked} in math mode. To include a bold face, lower case $\theta$, for example, we must use the construction

$$ \mbox{ \boldmath $\theta$ } $$

where the command \texttt{mbox}, which \textit{is} valid in math mode, temporarily shifts to text mode so the \texttt{boldmath} declaration can be invoked and the math mode expression in the argument of the command \texttt{mbox} will then be made bold; the bold presentation is confined by the command \texttt{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 ... \end}. Within the environment, each new item is introduced with the command \texttt{item}. Thus, the structure might look like

\footnote{For a complete list of all symbols and the commands that create them, see Section 3.3 on Mathematical Formulas in \textit{The \LaTeX\ Manual}.}
\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 `\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 `\item[DMC]` will cause the item to be labeled ‘DMC’ and the command `\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 `\item{}` to prevent the text in square brackets from being interpreted as the desired label.

- The length parameters `\itemsep`, `\parskip`, and `\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 `\labelitemi`, `\labelitemii`, `\labelitemiii`, and `\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 `\theenumi`, `\theenumii`, `\theenumiii`, and `\theenumiv`, whose action is in turn determined from the value of the corresponding one of the counters `enumi`, `enumii`, `enumiii`, and `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 `\Alph` could be replaced with `\alph`, `\roman`, `\Roman`, or `\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, and \LaTeX provides the \texttt{table} environment to facilitate proper placement of a table, 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. 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.

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. 

\footnote{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.}

\footnote{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.}

\footnote{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.}

\footnote{\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\{\}}\texttt{\}}, though we shall here not follow that recommendation. See \textit{The \LaTeX Manual} for details.}

\footnote{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.}
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.

A.7 Including Illustrations

Graphics displays frequently appear in technical documents. Within \LaTeX, means to incorporate graphics include:

- 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{\includegraphics} from the package \texttt{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. 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 \texttt{.aux} file so that the commands \texttt{\ref} and \texttt{\pageref} will function here as described in the first paragraph of Section A.4.

A.7.1 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{??\text{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,

\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.}

\footnote{The \texttt{figure} environment also admits the optional argument described in footnote 31.}
whichever position is adopted, the command \label must appear after the command \caption. For this method, you need only have printed copies of your 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 \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 graphicx package, whose features are made available by including that package with the statement

\usepackage{graphicx}

in the preamble of our \LaTeX code.

After placing the command \usepackage{graphicx} in our preamble, we incorporate the graphical display itself by placing a statement like

\includegraphics[height=dimension]{filename}

at the point where the PostScript or PDF figure defined in by the file filename is to be inserted. In general, the image would be centered horizontally on the page by inserting this command in a center environment, e.g.,

\begin{center}
  \includegraphics[height=dimension]{filename}
\end{center}

For example, the code

---

38 Appropriately translated, the substance of footnote 34 applies here to the entry placed automatically in the list of figures.

39 The 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 graphicx package, i.e., the extended graphics package.

40 More generally, this statement admits an optional argument to the desired graphics driver explicitly. Frequently, that driver will be dvips and the statement would be \usepackage[dvips]{graphicx}, but numerous other drivers exist, and many may be installed at your site. See the Local Guide for more information. For ultimate flexibility in the formatting of graphics files, we elect to leave that choice to the default.

41 If you use latex, dvips, and 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 \includegraphics command—the processing will correctly identify and incorporate the figures. If you seek to produce a PDF document directly with pdflatex as described in item 1 in Section A.1.3, then—even if the file type .pdf is omitted in the \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 vice versa is described in Section A.11. If both PostScript and PDF files exist for all figures and you omit the file type in the \includegraphics command, then you can use either latex or pdflatex for the processing.

42 Appropriately translated, the substance of footnote 34 applies here to the entry in the list of figures.
APPENDIX A. INTRODUCTION TO $\LaTeX$

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 diffract is stored in the current default directory. Here, the specification of the parameter height in the optional argument of the command \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 before the command \caption, so that the caption will appear above the figure. Further, the command \label, which must follow the command \caption, defines a symbolic label that can be used within the document to refer to the figure.

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 $\LaTeX$ document by the procedures described above, graphs produced by all toolkits will be properly incorporated in the .dvi file but those produced by the qt toolkit may not then translate with dvips to the subsequent PostScript file. When the latter is the objective, use gnuplot (or maybe fltk). A bit of testing may be necessary.

The command \includegraphics admits several optional arguments. The argument width can also be included. If width is specified instead of height, the height will be scaled to preserve

\textsuperscript{43}Omission of the file type in the statement including the graph prepares the way to use either latex-dvips to produce a PostScript file using latex and dvips (Section A.1.2), in which case the file diffract.ps or diffract.eps will be used, or pdflatex to produce a PDF file (Section A.1.3), in which case the file diffract.pdf will be used. The appropriate file must, of course, be accessible in either case.

\textsuperscript{44}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.
A.7. INCLUDING ILLUSTRATIONS

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 \texttt{\LaTeX} Graphics Companion.

A.7.3 Using the \texttt{tikz} Package

The \texttt{tikz} package, which is routinely included in present-day \texttt{\LaTeX} distributions, provides an assortment of \texttt{\LaTeX} macros to facilitate the drawing of simple—and even complicated—figures. Basically, the statement \texttt{\usepackage{tikz}} in the preamble followed in the document itself by statements embedded in a \texttt{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{Be aware that the code inserted in the \texttt{tikzpicture} environment describes the picture in the same way that \texttt{\LaTeX} provides the text and its formatting. Just as one imagines the final formatted document as the \texttt{\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 not a WYSIWYG 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 \texttt{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 \texttt{pgfmanual.pdf}. Further, googling \texttt{tikz} will bring up links to numerous documents, including in particular a link to \texttt{pgfmanual.pdf} and a link to the much more compact tutorial-style introduction in the file \texttt{minimaltikz.pdf}. Once you have selected any of these items, you may in some operating systems have to look in your \texttt{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 \texttt{\LaTeX} unit of length can be specifically stipulated. The optional argument \texttt{[x=1.0in,y=1.0in]} following \texttt{\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.

\footnote{Note that, when displayed on the screen with \texttt{xdvi} or \texttt{yap}—see Section A.1.4—the intermediate .\texttt{dvi} file produced in Section A.1.2 will not contain the correct figures.}
• **Every statement** must end with a semicolon. Any statement can be spread over several lines with only the last line so terminated.

• The statement `\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.\(^{48}\) Available colors include red, green, blue, cyan, magenta, yellow, and many others. You can also define your own color using a command like\(^{49}\)

\begin{verbatim}
\definecolor{ColorName}{rgb}{r,g,b}
\end{verbatim}

where *ColorName* is the name you assign to the color and *r*, *g*, *b* specify the *r*, *g*, and *b* components that make up the color. Available shapes include *rectangle*, *ellipse*, and *arc*.\(^{50}\)

• The statement `\draw [->, 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 `->` places an arrow at both ends of the line,\(^{51}\) and the argument *line width* species the weight of the line, by default, in points.\(^{52}\)

• Text is placed where specified with the statements `\node at (0,1.3) {\Large\bf y};` and `\node at (2.4,0.0) {\Large\bf x};`. Note that \LaTeX\ stipulations of font size and style are recognized by tikz. Note also that the text is, by default, *centered* at the specified point. See the TIKZ manuals for ways to override that default positioning.

• The statement `\draw [ultra thick, domain=0:2.0] plot (\x, {sin(pi*\x r)} );` illustrates how to draw a smooth curve defined by a function, many of which are available within tikz. Note also the availability of the irrational number π with the simple coding `pi`.\(^{53}\)

• The construction

\begin{verbatim}
\foreach \x in {0.0,1.0,2.0} {
 \draw (\x,0.1 )--(\x,-0.1);
 \node at (\x, -0.4){\x};
}
\end{verbatim}

creates a loop executing the statements enclosed in `{...}` for each value in the list following the keyword 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 `\draw` is a versatile command that admits numerous embellishments through the use of keywords and/or optional arguments. Conveniently, 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 tikz coding.

When features of the package 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

\begin{verbatim}
\foreach \x in {0.0,1.0,2.0} {
 \draw (\x,0.1 )--(\x,-0.1);
 \node at (\x, -0.4){\x};
}
\end{verbatim}

\(^{48}\)Note that the default unit—here inches—for radius can be overridden with an explicit stipulation, e.g., `[radius=0.1cm]`.

\(^{49}\)The command `\definecolor` here used is defined within the `tikzpicture` environment and does not require use of the \LaTeX\ package `color`.

\(^{50}\)See the manuals for the details of how to specify the dimensions of these shapes.

\(^{51}\)The separate arguments `->` and `<-` will place an arrow at the end or the beginning of the line, respectively.

\(^{52}\)The argument `line width=2` could be replaced, among others, by `ultra thin`, `very thin`, `thin`, `thick`, `very thick`, and `ultra thick`, and the default unit could be overridden with an argument like `line width=0.1cm`.

\(^{53}\)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 `\draw` command.
Figure A.2: An illustrative `tikz` figure. Note that specified colors will be converted to a gray scale unless the display device—screen or printer—can display colors.

```
\begin{figure}
  \caption{An illustrative \texttt{tikz} figure.}
  \label{LATEX:tikzfig}
  \begin{center}
  \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) {\Large\bf y};
    \node at (2.4,0.0) {\Large\bf 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}
```

`pdflatex filename` (Default file type `.tex`)

will produce a PDF file and the statements

`latex filename` (Default file type `.tex`)

`dvips -o filename.ps -t letter filename` (Default file type `.dvi`)

`ps2pdf filename.ps`

will produce a PostScript and then a PDF file, though figures defined by `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 A.11 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.

A.7.4 Using the picture Environment and 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.

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 .toc, .lof, and .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 .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 .lof or .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}}
Here, \textit{file} is one of \texttt{toc}, \texttt{lof}, and \texttt{lot}. \textit{unit} is one of \texttt{part}, \texttt{chapter}, \texttt{section}, \texttt{subsection}, or \texttt{subsubsection} if \textit{file} is \texttt{toc}, \texttt{figure} if \textit{file} is \texttt{lof}, and \texttt{table} if \textit{file} is \texttt{lot}.

These two commands have quite different effects. The first adds a \textit{bona fide} entry to the corresponding file. For example, the line

\begin{verbatim}
\addcontentsline{toc}{subsection}{Specially marked point}
\end{verbatim}

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

\begin{verbatim}
\addtocontents{toc}{\vspace{24pt}This is a note.\vspace{24pt}}
\end{verbatim}

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 \textit{The L\TeX\ Manual}. Note, in particular, the optional argument in the command \texttt{\caption} and the several sectioning commands (\texttt{\chapter}, \texttt{\section}, \ldots), which provide control over the entry each command places in the table of contents, list of figures, and list of tables.

\section*{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 \texttt{makeidx}, which is still available. We here recommend using \texttt{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.}

\begin{verbatim}
\usepackage{imakeidx}
\makeindex
\end{verbatim}

in the preamble and the command

\begin{verbatim}
\printindex
\end{verbatim}

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

\begin{verbatim}
\index{Item to be indexed}
\end{verbatim}

at appropriate points throughout the document. (The detailed structure of the argument to the command \texttt{index} is described in Section 4.5 and Appendix A of \textit{The \LaTeX\ Manual}.\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.} When the file \texttt{filename.tex} is processed by \texttt{latex} or \texttt{pdflatex}, the presence of these commands results in the addition of three auxiliary ASCII files named

\begin{itemize}
  \item \texttt{filename.idx} containing one line for each index entry,
  \item \texttt{filename.ilg} containing a log of messages created when, behind the scenes, the idx file is automatically processed through \texttt{makeindex}, and
\end{itemize}
• 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).\footnote{With \makeindex, 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.}

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 \TeX 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"\times9" area on 8.5"\times11" paper (\texttt{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 \texttt{ghostview} or acroread. Conveniently, if the .dvi file is displayed on the screen with \texttt{xdvi} (UNIX) or \texttt{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

\begin{verbatim}
\setlength{\textwidth}{6.0truein}
\setlength{\oddsidemargin}{0.5truein}
\setlength{\evensidemargin}{0.0truein}
\end{verbatim}

in the preamble with the commands

\begin{verbatim}
\setlength{\textwidth}{4.0truein}
\setlength{\oddsidemargin}{1.0truein}
\setlength{\evensidemargin}{1.0truein}
\end{verbatim}

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 \texttt{marginparwidth} (which sets the width of the marginal boxes) and \texttt{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.\footnote{For example, the \texttt{-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 \texttt{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 \texttt{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}.}

\begin{itemize}
\item [60] 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}) .aux, .idx, .ilg, .ind, .log, .out, .pdf, and .toc. The .idx, .ilg, and .ind files will be present only if you are creating an index with \texttt{imakeidx}. The .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.

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 \textit{above} the item rather than below the item.

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 \textit{same} source file both with \texttt{latex} and with \texttt{pdflatex}.

### 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 figure.eps and figure.ps to a useful PDF file involves the two steps

1. \texttt{ps2pdf figure.eps} \hspace{1em} (which will yield figure.eps.pdf)
   \texttt{ps2pdf figure.ps} \hspace{1em} (which will yield figure.pdf)

2. \texttt{pdfcrop figure.eps.pdf} \hspace{1em} (which will yield figure.eps-crop.pdf) or
   \texttt{pdfcrop figure.pdf} \hspace{1em} (which will yield 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}
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>rdfileeps (for ps files)</td>
<td>./rdfileeps (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 latex or 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
486

APPENDIX A. INTRODUCTION TO LATEX

pdftops FileName.pdf FileName.ps
pdftops -eps FileName.pdf FileName.eps
will effect this conversion. The task of creating scripts to automate this process when many files are
to be converted is left to the reader.

A.12

Using Conditional Expressions in LATEX

The LATEX package 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
\usepackage{ifthen}
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
\newboolean{FlagName}
\setboolean{FlagName}{FlagValue}
where FlagName can be any name that does not conflict with names already used and FlagValue
will be either true or false. Finally, at the point in the source file where some text is to be included
conditionally, we would place the command
\ifthenelse{\boolean{FlagName}}
{Text to be inserted if FlagValue is true.}
{Text to be inserted if FlagValue is false.}
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 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 CPSUP, the command
\usepackage{ifthen}
appears and several flags are defined and set with commands like, for example,
\newboolean{LATEX} \setboolean{LATEX}{true}
Finally, the command
\ifthenelse{\boolean{LATEX}}
{ \input{FileName} }
{ }
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 true and omitted if the flag is false.
More complicated logical operations can be constructed with the operators \and, \or, and \not.
For example, if the Boolean flags NUMREC and FORTRAN are defined and set, then the composite
statement


A.13. ERROR MESSAGES GENERATED BY \LaTeX

\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 latex or pdflatex and also to create easily either a printable document or a linked document, the package 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 ifthen can be omitted if it had been entered previously for other reasons.}

\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} }

which define the boolean variable PRINT, ask for entry of either true or false at execution time, and terminate execution if an invalid value is entered. Then, the lines

\usepackage{color}
\ifthenelse{\boolean{PRINT}}{}{\definecolor{MyPurple}{rgb}{0.577,0.000,1.000} \usepackage[colorlinks=true, linkcolor={MyPurple}]{hyperref}}

also in the preamble, will include the package hyperref, but only if PRINT is false.

One further conditional command included in latex and pdflatex without adding a package facilitates testing whether a file exists. For example, the statement

\IfFileExists{test.tex}{\input{test.tex}{}}

inputs 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 latex or 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.,

\IfFileExists{test.tex}{\input{test.tex}{\typeout{File not found} \end{document}}}

A.13 Error Messages Generated by \LaTeX

However carefully the code is created, sooner or later \LaTeX will generate an error message, which will have the general form

! error identification
1. number text read
text not read
?

\footnote{Inclusion of the package ifthen can be omitted if it had been entered previously for other reasons.}
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.

A.14 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

```
xdvi filename or yap filename
```

(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

\footnote{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.}

\footnote{Yet Another Previewer}

\footnote{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.}

\footnote{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 Button Action</th>
<th>Button</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>File→Reload</td>
<td>Reread</td>
<td>Rereads current \texttt{.dvi} file.</td>
</tr>
<tr>
<td>File→Quit</td>
<td>Quit</td>
<td>Quits previewer.</td>
</tr>
<tr>
<td>Zoom→Shrink by 1</td>
<td>100%</td>
<td>Displays text at full size (quite large).</td>
</tr>
<tr>
<td>Zoom→Shrink by 3</td>
<td>33%</td>
<td>Displays text at 33% of full size.</td>
</tr>
<tr>
<td>Zoom→Shrink by 4</td>
<td>25%</td>
<td>Displays text at 25% of full size.</td>
</tr>
<tr>
<td>Zoom→Shrink by 6</td>
<td>17%</td>
<td>Displays text at 17% of full size.</td>
</tr>
<tr>
<td>Navigate→First Page</td>
<td>First</td>
<td>Moves to first page.</td>
</tr>
<tr>
<td>Navigate→Page-10</td>
<td>Page-10</td>
<td>Moves to the tenth page before the one displayed.</td>
</tr>
<tr>
<td>Navigate→Page-5</td>
<td>Page-5</td>
<td>Moves to the fifth page before the one displayed.</td>
</tr>
<tr>
<td>Navigate→Prev</td>
<td>Prev</td>
<td>Moves to previous page.</td>
</tr>
<tr>
<td>Navigate→Next</td>
<td>Next</td>
<td>Moves to next page.</td>
</tr>
<tr>
<td>Navigate→Page+5</td>
<td>Page+5</td>
<td>Moves to the fifth page after the one displayed.</td>
</tr>
<tr>
<td>Navigate→Page+10</td>
<td>Page+10</td>
<td>Moves to the tenth page after the one displayed.</td>
</tr>
<tr>
<td>Navigate→Last Page</td>
<td>Last</td>
<td>Moves to last page.</td>
</tr>
<tr>
<td>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>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>s</td>
<td>Size text to fit window.</td>
</tr>
<tr>
<td>1s</td>
<td>(one-s) Largest text.</td>
</tr>
<tr>
<td>ns</td>
<td>Apply shrink factor \textit{n}.</td>
</tr>
<tr>
<td>g</td>
<td>Move to last page.</td>
</tr>
<tr>
<td>1g</td>
<td>(one-g) Move to first page.</td>
</tr>
<tr>
<td>ng</td>
<td>Move to (absolute) page \textit{n}.</td>
</tr>
<tr>
<td>n</td>
<td>Move to next page.</td>
</tr>
<tr>
<td>p</td>
<td>Move to previous page.</td>
</tr>
<tr>
<td>nn</td>
<td>Advance \textit{n} pages.</td>
</tr>
<tr>
<td>np</td>
<td>Back up \textit{n} pages.</td>
</tr>
<tr>
<td>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 \textit{Shell} window, leaving the \textit{Shell} window free for other uses. Then, with the \texttt{.tex} file in a text editor detached from the launching \textit{Shell} window, we can edit the text, save the edited version, use the \textit{Shell} window to invoke \texttt{\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.\footnote{In some environments, rereading an updated file will happen automatically without an explicit user request.} 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\footnote{To the author’s knowledge, no comparable stand-alone spell checker exists for Windows.}

The programs \texttt{ispell} and \texttt{aspell}, either or both of which may be part of your \texttt{\LaTeX} distribution, are common spell checkers. If either is installed at your site (see the \textit{Local Guide}), it can be invoked...
with a command like\textsuperscript{71}

\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 \LaTeX\ mode and will not identify every \TeX\ or \texttt{\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 \texttt{\LaTeX} is not assessed.

\subsection*{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
  \setlength{\textheight}{9truein} % default 10 point type
  \setlength{\topmargin}{-0.5truein} % Set height of text area
  \setlength{\textwidth}{6truein} % Center text on 11'' height
  \setlength{\oddsidemargin}{0.25truein} % Set width of text area
  \setlength{\evensidemargin}{0.25truein} % Center text on 8.5'' width
  \setlength{\parindent}{40pt} % Specify extra space
  \setlength{\parskip}{6pt plus 1pt minus 1pt} % between paragraphs,
                                          % allowing \LaTeX\ to adjust
  \setlength{\parindent}{0pt} % space 1 point up or down
  \setlength{\parindent}{0pt} % 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 \%
  \em Date\/: \today
\end{flushleft}
\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 \%
  \em Date\/: \today
\end{flushleft}
\end{document}
\end{verbatim}

\texttt{\LaTeX}’s default page setup to make fuller use of an 8.5''$\times$11'' page.\footnote{Note the special command $\backslash$\texttt{\LaTeX} provided for the display of the \LaTeX\ ‘‘logo’’.}

The title could have been generated using the three commands \texttt{\verb*+\title{+\ldots\verb*+}}, \texttt{\verb*+\author{+\ldots\verb*+}}, and \texttt{\verb*+\date{+\ldots\verb*+}} before the \texttt{\verb*+\begin{document}+}

\texttt{\LaTeX}’s default page setup to make fuller use of an 8.5''$\times$11'' page.\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{\texttt{man ispell}} or the command \texttt{\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+\texttt{?}+\verb+\texttt{?}+ 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_0y^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+.

\end{document} % Mandatory command

The output produced when this short sample file is processed through \LaTeX{} and \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 \emph{twice} by \LaTeX{}. The file itself is named \texttt{texsample2.tex} and can be copied from the directory \$\texttt{HEAD/tex}$.

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 \verb+\today+, which returns today’s date in the form “month date, year”.

4. The command \verb+\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}
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 a rather than 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 \setlength commands to modify \LaTeX's default page setup to make fuller use of an 8.5" × 11" page.\footnote{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 \verb|verbatim| environment.} Notice that a blank line creates a new paragraph. The \noindent command suppresses the standard paragraph indentation. In this special space we will also demonstrate some math environment operations. Consider the equation
\[
\int_0^\infty e^{-x} \, dx = 1,
\] (A.1)
where a medium sized space, specified by the command \texttt{\;}, is put between the integral sign \texttt{\int_0^\infty}, and the integrand and a small space, specified by the command \texttt{\,}, is inserted between the integrand and the \texttt{\,} in Eq. (A.1). The equation
\[
\frac{\partial e^{x_0 y^2}}{\partial x_0} = y^2 e^{x_0 y^2}
\] (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}.

\footnote{Note the special command \LaTeX provided for the display of the \LaTeX "logo".}

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
\[
\renewcommand{\today}{\number{day}\ space \ ifcase{month}or \ January% \ or \ February\ or \ March\ or \ April\ or \ May\ or \ June\ or \ July\ or \ August% \ or \ September\ or \ October\ or November\ or December\ fi \ space\number{year}}
\]

The percent signs at the end of the first two lines are not superfluous; they guarantee that \BbT\TeX will see this command as a single line and hence that extraneous spaces will not appear in

in the preamble will permit us to type the commands \texttt{\beq} and \texttt{\eeq} 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 \textit{The \BbT\TeX Manual}.
the output for occasional dates. Note also that this modification invokes \TeX’s case structure, which is introduced by the \texttt{\ifcase} command and terminated by the \texttt{\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 \LaTeX 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 \LaTeX Manual}.

8. \texttt{Re\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 \REVTex Input Guide} prepared by the AIP, the APS, and the OAS.\textsuperscript{72}

9. The command \texttt{\input}, which simply inputs the file specified in its argument. Details are described in Section 4.4 of \textit{The \LaTeX 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{\textbackslash cal} or the command \texttt{\mathcal}. It is described in Sections 3.3.2, 3.3.8, and C.7.8 of \textit{The \LaTeX Manual}.

11. Methods for placing accents over characters. For example, the command \texttt{\textasciitilde{o}} will produce \ö, the command \texttt{\grave{n}} will produce \={n}, and the command \texttt{\check{c}} will produce ¸c. A full listing of the possibilities will be found in Section 3.2.1 and Table 3.1 of \textit{The \LaTeX 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 \LaTeX document with the simple command sequence

\begin{verbatim}
\begin{verbatim} \input{FileName} \end{verbatim}
\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 not be recognized as introducing a command. If, instead, we invoke the alternative \texttt{alltt} environment with the \LaTeX commands

\begin{verbatim}
\begin{alltt} \input{FileName} \end{alltt}
\end{verbatim}

the problem is solved, since \texttt{\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 \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

\begin{verbatim}
\usepackage{alltt}
\end{verbatim}

placed in the preamble to the \LaTeX file.

\textsuperscript{72}Version 4.1, which was released in final form on 11 August 2010, is compatible with \LaTeX\textsuperscript{2}. The APS website \url{publish.aps.org/revtex4} (no www.) contains up-to-date information about Re\TeX\textsuperscript{4}, links to an assortment of manuals (including one titled \textit{Re\TeX\textsuperscript{4.1 Author’s Guide}}, and a link from which that version can be downloaded. Re\TeX\textsuperscript{4} is automatically included in many standard \LaTeX\textsuperscript{2} 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\ 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

\begin{verbatim}
\usepackage{verbatim}
\end{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.\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.}

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`c`, `\char`\`, and `{\tt \char`\`}` will produce the characters ‘b’, ‘c’, ‘‘’, 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 The \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
value when a new environment is entered. The remembered value must be stored in a \LaTeX variable called a \textit{counter}, which must be defined with the command

\begin{verbatim}
\newcounter{hold}
\end{verbatim}

where \texttt{hold}, which cannot contain numeric characters, is the name chosen for the counter. Then, the structure

\begin{verbatim}
\begin{enumerate}
\item Text of first item.
\item Text of second item.
\setcounter{hold}{\value{enumi}}
\end{enumerate}
\end{verbatim}

... Text not in enumerate environment.

\begin{verbatim}
\setcounter{enumi}{\value{hold}}
\begin{enumerate}
\item Text of third item.
\item Text of fourth item.
\setcounter{hold}{\value{enumi}}
\end{enumerate}
\end{verbatim}

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 \LaTeX 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 \LaTeX 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}
Table A.14: Structure to produce full-width text at top of double-column text.

\documentclass[twocolumn]{article}
  ... Whatever preamble we want
\begin{document}

\twocolumn
[ \begin{center}
  \{\large\bf Title\} \>[12pt] % Bold title; extra space
  \{\large\bf Author\} \>[12pt] % Bold author; extra space
\end{center}
\begin{quotation}
  \noindent ... Text of abstract.
\end{quotation}
]

... Text of paper
\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 \texttt{\the...}, where \ldots 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 \texttt{\twocolumn} command (not the optional argument \texttt{twocolumn} to the command \texttt{\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 \texttt{[12pt]} to put a bit of extra space after the title and author appears to generate error messages. The “fix” is to replace \texttt{[12pt]} with a full command, specifically \texttt{\vspace{12pt}}.}

If this pattern is to work, there can be no commands that produce output between the command \texttt{\begin{document}} and the command \texttt{\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 \texttt{\figurename}, but that portion of the label can be changed with a simple command like

\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 \texttt{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 \texttt{\documentclass} command at the beginning of the \LaTeX source file.\footnote{The file \texttt{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

$$\begin{bmatrix}
1 & 2 & 3 & 4 \\
5 & 6 & 7 & 8 \\
9 & 10 & 11 & 12
\end{bmatrix}$$

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 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 no multiplication sign after the number.
25. The Windows batch command `cleanTEX` defined by the batch file

```plaintext
REM delete all TEX intermediate files
```

and the Unix shell script defined by the file

```plaintext
#!/bin/bash
# delete all TEX intermediate files
rm -f *.sav *.out *.ind
```

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

```plaintext
REM Delete .bak files
del /s *.bak
del /s *.sav
```

and the UNIX shell script `cleanbak` defined by the the file

```plaintext
#!/bin/bash
# Delete *.bak, *.sav, and *~ files
find . -name '*.bak' -exec rm {} \;
find . -name '*.sav' -exec rm {} \;
find . -name '*~' -exec rm {} \;
```

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.

### 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 [www.tug.org](http://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

---

\(^{77}\)Files with file types `.aux`, `.log`, `.dvi`, `.toc`, `.ilg`, `.idx`, `.bak`, `.sav`, `.out`, and `.ind.`

\(^{78}\)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:


*Math Into \LaTeX* (Third Edition), George Grätzer (Birkäuser, Boston) [ISBN 0-8176-4131-9 or 978-0201430111, 2000]

*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.

### A.19 Exercises

**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.

**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
• your name, spaced far enough below the closing to allow for your signature and aligned with the closing.

Suppress page numbers altogether on this letter. Include both your \LaTeX 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 \LaTeX, explain the function of each command it contains. Include both your \LaTeX code and the processed output in what you submit as a solution to this exercise.

A.4. The folder $\$HEAD/tex$ 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

$$\frac{dA}{dt} = -k_AA; \quad \frac{dB}{dt} = k_AA - k_BB; \quad \frac{dC}{dt} = k_BB$$

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 \LaTeX code and the processed output in what you submit as a solution to this exercise. A Crutch: The file $\$HEAD/tex/sampledoc.tex$ provides a start on the creation of a suitable source file for this exercise.

A.5. Generate \LaTeX code to produce the memo

\begin{center}
\textbf{MEMORANDUM}
\end{center}

\textit{Current date}

\textbf{TO:} Insert name of recipient here.

\textbf{FROM:} Insert name of sender here.

\textbf{SUBJECT:} 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 \LaTeX 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\texttt{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 501. Create a \LaTeX source file that you can save and use repeatedly to facilitate generating each required report, i.e., create a \LaTeX 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:

<table>
<thead>
<tr>
<th>Category</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salaries</td>
<td>$xxx,xxx</td>
</tr>
<tr>
<td>Fringe benefits</td>
<td>$xxx,xxx</td>
</tr>
<tr>
<td>Equipment</td>
<td>$xxx,xxx</td>
</tr>
<tr>
<td>Supplies</td>
<td>$xxx,xxx</td>
</tr>
<tr>
<td>Travel</td>
<td>$xxx,xxx</td>
</tr>
<tr>
<td>Page charges</td>
<td>$xxx,xxx</td>
</tr>
<tr>
<td>Other: Identify</td>
<td>$xxx,xxx</td>
</tr>
</tbody>
</table>

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 ceps2pdf.bat

REM ceps2pdf.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
pdfcrop 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
pdfcrop 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
    pdfcrop 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
    pdfcrop 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: geospatial@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. 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 and computing.llnl.gov/projects/odepack/software. Additional information and downloads may be found from links at www.netlib.org/odepack. In particular, the two text files opkd-sum and
opiks-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 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 http://maxima.sourceforge.net/. MAXIMA is freely available for a wide variety of platforms.

**MAPLE\textsuperscript{®}**

Waterloo Maple, Inc.
615 Kumpf Drive
Waterloo, Ontario
CANADA, N2V 1K8

Voice: 800-267-6583
FAX: 519-747-5284
E-mail: info@maplesoft.com
Web: www.maplesoft.com

**MATHEMATICA\textsuperscript{®}**

Wolfram Research, Inc.
100 Trade Center Drive
Champaign, IL 61820-7237 USA

Voice: 217-398-0700, 800-965-3726
FAX: 217-398-0747
E-mail: info@wolfram.com
Web: www.wolfram.com

**MATLAB\textsuperscript{®}**

The MathWorks, Inc.
3 Apple Hill Drive
Natick, MA 01760-2098 USA

Voice: 508-647-7000
FAX: 508-647-7101
E-mail: info@mathworks.com
Web: www.mathworks.com

**MiK\TeX{}**

MiK\TeX{} is an implementation of the \TeX{}/\LaTeX{} system for Windows computers. Information is available at the URL miktex.org. MiK\TeX{} is also available from CTAN, the \TeX{}/\LaTeX{} distribution network accessible from links at 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 www2.cisl.ucar.edu/research-software/software. MUDPACK can be downloaded from the webpage 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.

Numerical Algorithms Group, Inc.
801 Warrenville Road
Suite 185
Lisle, IL 60532-4332 USA

Voice: 630-971-2337
FAX: 630-971-2706
E-mail: infodesk@nag.com
Web: www.nag.com
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/L\LaTeX{} system for Macintosh computers. Information is available at the URL www.trevorrow.com/oztex. OzTEX is also available from CTAN, the \TeX/L\LaTeX{} 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/L\LaTeX{}/dvips/graphicx/makeindex/xdvi/dvips/...

The primary site for information (history, current plans, downloads, ...) for \TeX/L\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/L\LaTeX{} system can be downloaded.

\TeX{}live

\TeX{}live is among the newer implementations of the \TeX/L\LaTeX{} system for all platforms. Information about this distribution and instructions for downloading and installing it are available from the URL www.tug.org/texlive.
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, 37, 38, 121, 222, 243, 267, 283, 321, 322, 370, 457

$\text{IDLHEAD}$, 11, 322

$\text{MATLABHEAD}$, 79, 322

$\text{MAXHEAD}$, 95

$\text{NRHEAD}$, 11

A

aborting execution
- in MATLAB, 16
- in MAXIMA, 94
accuracy
- in evaluating integrals, 278–279, 283–286
- in finding roots, 320–321
- in solutions via FEM, 431–433
- in solving ODEs, 219–221, 229, 232
  - with conservation laws, 221, 232–234
- in solving PDEs, 398–400, 405–407, 412–413

acoread, 462

action statements, 153, 156

Adams/Bashforth/Moulton (ABM) methods,
- 222, 227

adaptive methods
- for integrals, 280
- for ODEs, 221

adding text to graphics, see placing text in graphics

algorithms
- definition, 153
- for $a^2 + b^2 = c^2$, 179
- for adding, 179
- for bowling, 177
- for controlling loops, 162
- for counting, 162, 163
- for exchange, 162
- for finding extremes, 163, 164
- for integration, 273–283
- for Laplace’s equation, 164

for ODEs, 213–222
- for parking lot, 177
- for root finding, 318–321
- for searching, 164, 177
- for sorting, 179
- for summing, 163

animation
- in MAXIMA, 133

arithmetic operators
- in MATLAB, 22
- in MAXIMA, 98

array indices
- in MATLAB, 19, 170
- in MAXIMA, 114
- in pseudocode, 155

arrays
- for data storage, 166
- in MATLAB, 19–21
- in MAXIMA, 97, 113–115
- in pseudocode, 155

scalar, see scalar fields

vector, see vector fields

ASCII character codes, 5

ASCII text files, 8

aspell, 489–490

assigning values to variables
- in MATLAB, 17–18
- in MAXIMA, 99–100
- in pseudocode, 156

B

backsubstitution, 337

backward difference, 366

Bessel functions, 305, 344

binary files, 8

binary number system, 4–5

bisection, see root finding

bit, 4–5

black body radiation, 145, 302
blocking statements, see also control structures
in MATLAB, 33
in MAXIMA, 118, 332
in pseudocode, 158, 168

Boolean values, see logical expressions
Boolean variables, see logical expressions

boundary conditions
Dirichlet, 351, 365, 367–368
mixed, 351, 365, 367–368
Neumann, 351, 365
of first kind, see boundary conditions, Dirichlet
of second kind, see boundary conditions, Neumann
of third kind, see boundary conditions, mixed

boundary value problems, 187, 198, 347, 348, 351
byte, 5–6

C

C, 1
cardioid, 45, 299
carrying capacity, 192, 204–206
Cartesian coordinates
in MAXIMA, 150
case sensitivity
in MATLAB, 16
in MAXIMA, 94
in pseudocode, 154
center of mass
definition of, 256
of cardioid, 299
of semicircular plate, 257, 265
central difference, 366
central processing unit, 3
chaos
Lorenz system, 251
character codes, 4–5
character strings
in MATLAB, 17, 18
characteristic equation, 308
charging capacitor, 85
classification
of boundary conditions, 351–352
of integrals, 255
of ODEs, 187
of PDEs, 351
color
in MATLAB, see also MATLAB properties:
Color, ColorMap, ColorOrder, EdgeColor, FixedColors, 27, 76–77
combining statements on one line
in MATLAB, 16, 33, 46
in MAXIMA, 94
in pseudocode, 158
command files
in MATLAB, see MATLAB scripts
in MAXIMA, see MAXIMA command files
command line editing
in MATLAB, 16
in MAXIMA, 100
command/function duality
in MATLAB, 19, 21
commands
in MATLAB, see MATLAB commands
in MAXIMA, see MAXIMA commands
in pseudocode, see pseudocommands
comments, see also documentation
in \LaTeX, 490
in MATLAB, 16, 37, 40
in MAXIMA, 121
in pseudocode, 159, 162
common storage
in MATLAB, see global variables
compiler directives, 155
in pseudocode, see pseudocode directives
compiling source file, see Local Guide, 154
complex numbers
in MATLAB, 19
in MAXIMA, 104–105
Compton scattering, 86
conditions, see logical conditions
conservation of mass, 360–361
constants
in MATLAB, see MATLAB system variables
in MAXIMA, see MAXIMA special symbols
contact information, 505–508
continuing statement on second line
in MATLAB, 16
in MAXIMA, 94
in pseudocode, 158
control characters, 4
control structures
in MATLAB, see MATLAB control structures
in MAXIMA, see MAXIMA control structures
in pseudocode, see pseudocode control structures
convergence criteria, 166, 219–221, 278–279, 320–321
convergence error, 368–369, 395
convert .eps/.ps to .pdf, 484–485
convert .pdf to .eps/.ps, 485
Cornu spiral, 260, 290–292
coupled oscillators, 194–195, 208–209, 246
CPU, see central processing unit
Cramer’s rule, 336–337
cross product
in MATLAB, 83
in MAXIMA, 121–122, 152
current directory, see also default directory
in MATLAB, 30–32, 46, 78, 228, 287
curve fitting, see least squares fitting
to parabola, 149, 346
customizing
INDEX

MATLAB, 79
MAXIMA, 139
cylindrical coordinates
in MAXIMA, 151

D
data analysis, 259
data files
  freefall.dat, 346
  irrad.matlab.dat, 171, 176
  ldrum.dat, 69
  pdens.matlab.dat, 172
  quadpole.matlab.dat, 176
  radio.dat, 34, 35, 37, 80
  rcdata.dat, 46
  ruchardt.dat, 348
  wavegd.matlab.dat, 175
data types
  floating point, 6–7
  integer, 6–7
  string, 6–7
data typing
  explicit, 155
  implicit, 155
  in MATLAB, 17
  in pseudocode, 154–155, 157–158
default directory, see also current directory, 9
  in MATLAB, 235
  in MAXIMA, 137–138
defining ODEs
  for MATLAB, 226
  for MAXIMA, 110, 202, 204, 207, 209, 211, 244
deleting functions
  in MAXIMA, 141
deleting variables
  in MATLAB, 20, 33
  in MAXIMA, 100, 107, 141
demos
  in MATLAB, 16
  in MAXIMA, 96
diffraction patterns, see also interference
  patterns, 170, 304, 344, 345
  of square aperture, 50–58
diffusion equation
  in 1D, 356–357
    in MATLAB, 400–407
  in 2D, 389–392
  in 3D, 357
dimensionless casting
  for plotting, 87
  of integrals, 257–259, 261, 267, 298, 302
direct algorithm, 368
discretization
  of Laplace’s equation, 165
  of ODEs, 366–368
  of PDEs, 352, 373
discretization error, 368–369, 373, 395
displaying data on screen
  in MATLAB, 16, 19–21
  in MAXIMA, 94
  in pseudocode, 156
dividing statement between lines, see continuing
  statement on second line
documentation, see also comments
  in MATLAB M-files, 40
  in MAXIMA workbooks, 133–134
dot product
  in MATLAB, 23, 83
  in MAXIMA, 121–122, 152
double factorial, 12
drawing space curves
  with MATLAB, 73–74
  with MAXIMA, 140
dvipdfm, 462
dvips, 461, 479
dynamic allocation of memory, 155
E
elastic swing, 252
electric field, 260
  in wave guide, 65–68, 172–175
  of dipole, 182
  of quadrupole, 175–176
  of relativistic particle, 86
electrostatic potential, 260
  of charged disks, 85, 86, 299
  of charged rings, 260–261, 293, 297, 303
  of finite line charge, 266–267
  of spherically symmetric distribution, 299
elemental equation in finite element analysis, 378
elliptic integrals, 149, 259, 303
em dash, 470
em space, 467
en dash, 470
en space, 467
energy levels
  in quantum oscillator, see harmonic
  oscillator
  in quantum well, 210–211, 311, 327–329, 333–335, 344
equation of continuity, 360–361
equilibrium points, 306, 314
equivalent resistance, 346
error function, 260, 262, 283, 287, 289, 290, 296, 302
error messages
  in LATEX, 487–488
Euler’s method, 214–215, 223–225
  improved, see improved Euler method
evaluating integrals
  with MATLAB, 283–294
examples, see demos
exchange algorithm, 162
executing program, see Local Guide
exiting
  from MATLAB, 16
  from MAXIMA, 94
expressing infinity
  in MAXIMA, 264
expressions
  in MAXIMA, 97–99

F

Fabry-Perot interferometer, 85
factorial, 12, 101
double, 101
Fibonacci numbers, 83
field mapping, 89, 90
finding eigenvalues/eigenvectors
  with MATLAB, 41–42, 84
  with MAXIMA, 114–115, 147
finding roots, see root finding
finite difference approximations
  for derivatives, 179
  for PDEs, 352
finite difference methods
  error in, 368–369
  in 1D, 366–373
  in 2D, 387–413
  with MATLAB, 369–373, 395–413
finite element approximations
  for PDEs, 352
finite element methods
  assembling equations, 373
  in 1D, 373–387
  in 2D, 413–433
  incorporating boundary conditions, 373
  with MATLAB, 382–387, 423–433
fitting data, see curve fitting, least squares fitting
floating point number
  double precision, 7
  single precision, 6
fluid dynamics, 359
fonts, see also Greek letters
  in LATEX, 71–73
  in MATLAB, 166–167
formatting output
  in MATLAB, 30–31
  in pseudocode, 156
FORTRAN, 1
forward difference, 366
four-slit interference pattern, 87
Fourier series, 263, 272–273, 300
Fraunhofer diffraction, 170–171
Fresnel integrals, see also Cornu spiral, 260, 304
functions
  in MATLAB, see MATLAB functions
  in MAXIMA, see MAXIMA functions
typical available, 156

G

Galerkin method, 373, 375–378
gamma function, 12–13
Gauss-Jordan elimination, 373
Gaussian distribution, 259, 302
Gaussian elimination, 337–338, 373
Gaussian integration, see quadrature
gedit, 1
ghostview, 462
global positioning, 345
global truncation error, see truncation error, global
global variables
  in MATLAB, 39–40, 292
gnuplot command
  set, 127–130
gnuplot keywords
  border, 129–130
cntrparam, 127–130
cycle, 129–130
discrete, 128
incremental, 128
lc, 129–130
levels, 127–130
linestyle, 129–130
linetype, 129–130
linewidth, 129–130
lw, 129–130
ratio, 128–130
size, 128–130
golden ratio, 343
gravitational force, 196
Greek letters, see also fonts
  in LATEX, 470–471
  in MATLAB, 52, 72
  in MAXIMA, 97

H

handles
  in MATLAB, 29–30, 55, 64, 72, 74–75
hard copy
  from MATLAB, 49–50
  from MAXIMA, 130–132
Helmholtz coil, 267–271
Helmholtz equation, 358
help, see on-line help
hexadecimal number system, 5
high-pass filter, 46–49
high-resolution graphing
  in MATLAB, 25–27, 42–68
  in MAXIMA, 123–130
Hooke’s law, 307
hyperlinks, 483–484

I
IDL, 1
IDL contact information, 505
improved Euler method, 216, 224
indices in \LaTeX, 481
indicial equation, 211
inert commands
  in MAXIMA, see suppressing evaluation in MAXIMA
inhomogeneities in linear equations, 336
initial value problems, 187, 351
initialization files
  in MATLAB, 79
  in MAXIMA, 139
integer
  signed, 6
  unsigned, 6
interference patterns, see also diffraction patterns
  Fabry-Perot, 85
  from four slits, 87
internal storage of expressions
  in MAXIMA, 98–99
interpolation functions, see shape functions
  interpolation points, 276, 277
interpreting source code, see also Local Guide, 154
interrupt
  in MATLAB, 16
  in MAXIMA, 94
ispell, 489–490
iterative algorithm, 369
K
Kaprekar Constant, 91
keywords
  in MATLAB, see MATLAB properties
Kirchhoff's laws, 346
Knuth, Donald, 457
Kronecker delta, 262
L
Lagrange equations
  for 2D harmonic oscillator, 148
  for compound pendulum, 148
Laguerre polynomials, 146
Laguerre's method, see root finding
Lamport, Leslie, 457
LAPACK, 1
  contact information, 505
Laplace transforms, 13–14, 199–200, 248
  definition, 13, 109, 199
  for ODEs, 199
  in MAXIMA, 109–110, 149
Laplace's equation, 348, 357, 358
  in 2D, 392–395
  in MATLAB, 407–413
  in MATLAB, 169–170
  in pseudocode, 164–166
\LaTeX, 1
  accents in, 493
  counters in, 494
  dashes in, 469
  math mode in, 470–471
  quotation marks in, 470
  special characters in, 494
latex, 460, 461, 479
\LaTeX commands
  \textbackslash (italic correction), 466
  \textbackslash (new line), 468, 473
  \textbackslash ! (negative thin space), 471
  \textbackslash , (thin space), 471
  \textbackslash {-} (allow hyphen), 470
  \textbackslash :) (medium space), 471
  \textbackslash ; (thick space), 471
  \addcontentsline, 480–481
  \addtocounter, 480–481
  \Alph, 472, 495
  \alph, 472, 495
  \and, 486
  \arabic, 472
  \author, 490
  \bf, 490
  \bfseries, 466
  \cal, 493
  \caption, 473, 474, 476, 479
  \centerline, 476
  \chapter, 470
  \chaptername, 497
  \char, 494
  \cleardoublepage, 467
  \contentsname, 497
  \date, 490
  \definecolor, 483, 487
  \documentclass, 458, 460, 463, 484, 490
  \documentstyle, 458
  \fbox, 494
  \figurename, 496
  \footnote, 470, 490
  \footnotesize, 467
  \frac, 490
  \framebox, 494
  \hskip, 467, 468
  \hspace, 467
  \Huge, 467
  \huge, 467
  \IFDEFDEF, 487
  \ifthenelse, 486–487
  \includegraphics, 474–477, 484
  \index, 481
  \indexname, 497
  \input, 493
  \int, 490
  \item, 471–472, 495
  \itshape, 466
  \label, 470, 473, 474, 476, 479, 490
<table>
<thead>
<tr>
<th>Command</th>
<th>Page(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>\labelitemi</td>
<td>472</td>
</tr>
<tr>
<td>\labelitemii</td>
<td>472</td>
</tr>
<tr>
<td>\labelitemiii</td>
<td>472</td>
</tr>
<tr>
<td>\labelitemiv</td>
<td>472</td>
</tr>
<tr>
<td>\LARGE</td>
<td>467</td>
</tr>
<tr>
<td>\Large</td>
<td>467, 490</td>
</tr>
<tr>
<td>\large</td>
<td>467</td>
</tr>
<tr>
<td>\LaTeX</td>
<td>490</td>
</tr>
<tr>
<td>\listoffigures</td>
<td>480–481</td>
</tr>
<tr>
<td>\listoftables</td>
<td>480–481</td>
</tr>
<tr>
<td>\makebox</td>
<td>494</td>
</tr>
<tr>
<td>\makeindex</td>
<td>481–482</td>
</tr>
<tr>
<td>\maketitle</td>
<td>490</td>
</tr>
<tr>
<td>\mathbf</td>
<td>471</td>
</tr>
<tr>
<td>\mathcal</td>
<td>493</td>
</tr>
<tr>
<td>\mathit</td>
<td>471</td>
</tr>
<tr>
<td>\mathrm</td>
<td>471</td>
</tr>
<tr>
<td>\mathstt</td>
<td>471</td>
</tr>
<tr>
<td>\mbox</td>
<td>494</td>
</tr>
<tr>
<td>\makeindex</td>
<td>481–482</td>
</tr>
<tr>
<td>\maketitle</td>
<td>490</td>
</tr>
<tr>
<td>\mathbf</td>
<td>471</td>
</tr>
<tr>
<td>\mathcal</td>
<td>469</td>
</tr>
<tr>
<td>\mathit</td>
<td>471</td>
</tr>
<tr>
<td>\mathrm</td>
<td>471</td>
</tr>
<tr>
<td>\mathstt</td>
<td>471</td>
</tr>
<tr>
<td>\mbox</td>
<td>470, 471, 494</td>
</tr>
<tr>
<td>\mdseries</td>
<td>466</td>
</tr>
<tr>
<td>\Needspace</td>
<td>497</td>
</tr>
<tr>
<td>\needspace</td>
<td>497</td>
</tr>
<tr>
<td>\newboolean</td>
<td>486–487</td>
</tr>
<tr>
<td>\newcommand</td>
<td>491</td>
</tr>
<tr>
<td>\newcounter</td>
<td>495</td>
</tr>
<tr>
<td>\newlength</td>
<td>468</td>
</tr>
<tr>
<td>\newpage</td>
<td>467</td>
</tr>
<tr>
<td>\noindent</td>
<td>467, 490</td>
</tr>
<tr>
<td>\normalfont</td>
<td>495</td>
</tr>
<tr>
<td>\normalfont</td>
<td>495</td>
</tr>
<tr>
<td>\not</td>
<td>486</td>
</tr>
<tr>
<td>\or</td>
<td>486</td>
</tr>
<tr>
<td>\pageref</td>
<td>470, 473, 474</td>
</tr>
<tr>
<td>\pagestyle</td>
<td>490, 493</td>
</tr>
<tr>
<td>\parbox</td>
<td>494</td>
</tr>
<tr>
<td>\part</td>
<td>470</td>
</tr>
<tr>
<td>\partial</td>
<td>490</td>
</tr>
<tr>
<td>\printindex</td>
<td>481</td>
</tr>
<tr>
<td>\quad</td>
<td>467</td>
</tr>
<tr>
<td>\quad</td>
<td>467</td>
</tr>
<tr>
<td>\ref</td>
<td>470, 473, 474, 490</td>
</tr>
<tr>
<td>\renewcommand</td>
<td>469, 472, 492, 495, 496</td>
</tr>
<tr>
<td>\rmfamily</td>
<td>466</td>
</tr>
<tr>
<td>\Roman</td>
<td>472, 496</td>
</tr>
<tr>
<td>\roman</td>
<td>472, 496</td>
</tr>
<tr>
<td>\rule</td>
<td>494</td>
</tr>
<tr>
<td>\scriptsize</td>
<td>467</td>
</tr>
<tr>
<td>\scshape</td>
<td>466</td>
</tr>
<tr>
<td>\section</td>
<td>470</td>
</tr>
<tr>
<td>\setboolean</td>
<td>486–487</td>
</tr>
<tr>
<td>\setcounter</td>
<td>495</td>
</tr>
<tr>
<td>\setlength</td>
<td>463, 467, 490</td>
</tr>
<tr>
<td>\settodepth</td>
<td>467</td>
</tr>
<tr>
<td>\settoheight</td>
<td>467</td>
</tr>
<tr>
<td>\settowidth</td>
<td>468</td>
</tr>
<tr>
<td>\ssfamity</td>
<td>466</td>
</tr>
<tr>
<td>\ssshape</td>
<td>466</td>
</tr>
<tr>
<td>\small</td>
<td>467</td>
</tr>
<tr>
<td>\subsection</td>
<td>470</td>
</tr>
<tr>
<td>\thesubsection</td>
<td>470</td>
</tr>
<tr>
<td>\tablename</td>
<td>497</td>
</tr>
<tr>
<td>\tableofcontents</td>
<td>480–481</td>
</tr>
<tr>
<td>\textbf</td>
<td>466</td>
</tr>
<tr>
<td>\textit</td>
<td>466</td>
</tr>
<tr>
<td>\textmd</td>
<td>466</td>
</tr>
<tr>
<td>\textnormal</td>
<td>466</td>
</tr>
<tr>
<td>\textsf</td>
<td>466</td>
</tr>
<tr>
<td>\texttt</td>
<td>466</td>
</tr>
<tr>
<td>\textup</td>
<td>466</td>
</tr>
<tr>
<td>\the...</td>
<td>495–496</td>
</tr>
<tr>
<td>\thenumii</td>
<td>472</td>
</tr>
<tr>
<td>\thenumiii</td>
<td>472</td>
</tr>
<tr>
<td>\thenumiv</td>
<td>472</td>
</tr>
<tr>
<td>\thispagestyle</td>
<td>490, 493</td>
</tr>
<tr>
<td>\tiny</td>
<td>467</td>
</tr>
<tr>
<td>\title</td>
<td>490</td>
</tr>
<tr>
<td>\today</td>
<td>491, 492</td>
</tr>
<tr>
<td>\ttfamily</td>
<td>466</td>
</tr>
<tr>
<td>\twocolumn</td>
<td>496</td>
</tr>
<tr>
<td>\typein</td>
<td>487</td>
</tr>
<tr>
<td>\typeout</td>
<td>487</td>
</tr>
<tr>
<td>\upshape</td>
<td>466</td>
</tr>
<tr>
<td>\usepackage</td>
<td>469, 475, 481</td>
</tr>
<tr>
<td>\value</td>
<td>495</td>
</tr>
<tr>
<td>\verb</td>
<td>495</td>
</tr>
<tr>
<td>\verbatiminput</td>
<td>494</td>
</tr>
<tr>
<td>\vspace</td>
<td>467, 474</td>
</tr>
<tr>
<td>\vspace*</td>
<td>467</td>
</tr>
</tbody>
</table>

**LaTeX** contact information, 457, 507

**LaTeX** counters

- enumi, 472, 495
- enumii, 472
- enumiii, 472
- enumiv, 472
- figure, 495
- footnote, 495
- page, 495
- secmnumdepth, 480
- section, 495
- subsection, 495
- subsubsection, 495
- table, 495
- tocdepth, 480

**LaTeX** declarations, see **LaTeX** commands

**LaTeX** document classes, 459

- article, 463, 470
- book, 463, 470
- report, 463
- slides, 463

**LaTeX** environments, 468–469

- \(...\) (in-line equation), 470
- \[...\] (displayed equation), 470
- $...$ (in-line equation), 470
- array, 497
INDEX

center, 468, 473, 475, 479, 490
comment, 494
displaymath, 470
document, 459, 460, 490
enumerate, 471–472, 494, 495
eqarray, 470
equation, 470, 490
figure, 474–477, 479
flushleft, 490
itemize, 471–472
minipage, 494
pict2e, 480
picture, 474, 480
quotation, 468
quote, 468
table, 472–474	
tabular, 472–474
tikzpicture, 477–480

\LaTeX\ length parameters, 464
\baselineskip, 467, 469
\baselinestretch, 469
\bottomfraction, 495
\evensidemargin, 463, 482
\floatpagefraction, 495
\itemsep, 472
\marginparsep, 482
\marginparwidth, 482
\oddsidemargin, 463, 490
\parsep, 472
\parskip, 463, 472, 490
\textfraction, 495
\textheight, 463, 490
\textwidth, 463, 482, 490
\topfraction, 495
\topmargin, 463, 490

\LaTeX\ Manual, 457
\LaTeX\ output
from MAXIMA, 132–133
\LaTeX\ packages, 469
alltt, 493
color, 483, 487
graphicx, 474–477, 483–484
hyperref, 483–484, 487
ifthen, 486–487
imakeidx, 481–482
makeidx, 481–482
needspace, 497
showidx, 482
tikz, 474, 477–480
verbatim, 493–494

\LaTeX\ page previewers, see also xdvi, yap, 488–489
\LaTeX\ preamble, 463–465
\LaTeX\ sample, 459–462, 490–491
\LaTeX\ special symbols
" " (opening double quotation), 470
--- (en dash), 469
-- (en dash), 469

- (dash), 469
- (minus sign), 470
" (hard space), 469
" " (closing double quotation), 470
\LaTeX\ templates, 464
\LaTeX\ type size, 466
\LaTeX\ type styles, 465–466
least squares fitting, 346, 348
Legendre polynomials, 143, 144, 282, 300, 343
Lennard-Jones potential, 345
licensing of software, 14
line labels
in MAXIMA, 94
linking object code, see Local Guide, 154
lists
in MAXIMA, 97–99, 113–115
Local Guide, 2, 8, 9, 11, 14, 15, 37, 50, 78, 93, 94, 121, 170, 222, 243, 267, 283, 321, 457, 459–463, 465, 475, 488, 490
logical conditions, see logical expressions
logical expressions
in \LaTeX, 486–487
in MATLAB, 33–34
in MAXIMA, 119
in pseudocode, 157–158
logical operators, see logical expressions
logical unit number (LUN)
in MATLAB, 31
logistic growth, 192, 204–206
long integer
signed, 6
unsigned, 6
loops
in MATLAB, 32–34
in MAXIMA, 102, 106, 107, 117–119
in pseudocode, 159–161
Lorentz distribution, 85, 298, 302
Lorentz lineshape, 38, 39
lp, see UNIX commands
LSODE, 1
LSODE contact information, 505
LU decomposition, 338, 373

M

M-files, see MATLAB scripts, MATLAB functions
machine code, 154
Macintosh OS, 1
Mac\TeX
contact information, 506
magnetic field, 260
in wave guide, 65–68, 90, 172–175
of circular loops, 43, 87, 181, 267, 303
of Helmholtz coil, 267–271
of solenoid, 87
magnetic vector potential, 260
MAPLE, 1
INDEX

MAPLE contact information, 506
MATHEMATICA, 1
MATHEMATICA contact information, 506
mathematical functions
  in MATLAB, see also MATLAB functions,
  MATLAB programs, MATLAB commands, 22
  in MAXIMA, see MAXIMA functions
mathematical symbols
  in MATLAB, 72
MATLAB, 1
matlab, 15
MATLAB attributes
  complex, 20
MATLAB command files, see MATLAB scripts
MATLAB commands, see also MATLAB
  functions, MATLAB programs
  addpath, 78
  axes, 28, 29, 71
  axis, 26, 27, 47, 55, 62–64, 69, 70, 75–76,
    238, 239, 241, 291, 294
  camlight, 64
  cd, 30, 78
  clabel, 55
  clear, 20, 33, 284
  clf, 64, 74–75
  close, 74
  colorbar, 63, 64
  colormap, 56, 64, 77
  common, 232
  contour, 51, 54–55, 59, 62, 68
  contourf, 55
  cross, 23
  cumsum, 83
  demo, 16, 37
  det, 23
  diary, 31–32, 49
  disp, 81, 395
  dot, 23
  double, 17
  eig, 84
  errorbar, 47–48
  fclose, 30, 31, 35, 37, 46, 69, 80
  feval, 38
  fgetl, 35, 37, 80
  figure, 28, 29, 74–75
  fix, 17, 22, 382, 397, 398, 408, 410, 424
  fopen, 30, 31, 35, 37, 46, 69, 80
  for matrix multiplication, 23
  format, 21, 79, 283, 287, 327
  fprintf, 30, 31, 81, 408, 410
  fscanf, 31, 35, 37, 46, 69, 80
  fzero, 321–323, 326
  gca, 29, 30, 35, 43, 45, 62–64, 72, 79, 223,
    228, 235, 324–326, 329, 396–397, 400,
    402–403, 405–406, 431
  gcf, 29, 30, 74–75, 77
  general structure of, 18–19
  get, 29, 64, 72, 74–75, 77, 79
  getframe, 69, 70
  global, 40
  grid, 26–28, 73, 235, 290, 291, 294,
    324–326, 371, 385
  help, 22, 36–37, 40, 50, 243
  helpdesk, 37
  helpwin, 37
  hold, 35, 37, 44, 47–48, 55, 64, 68, 75, 223,
    228, 230, 234, 239, 241, 324, 325, 329,
    371
  input, 81, 369, 382, 395, 397, 400, 403, 408,
    412, 424
  int16, 17
  int2str, 17
  int32, 17
  int64, 17
  int8, 17
  inv, 23, 24
  isosurface, 64
  length, 293
  lighting, 64
  line, 28, 29
  linspace, 25, 286, 289, 290
  load, 30
  loglog, 47
  logspace, 48
  lookfor, 40
  lu, 89
  lubisect, 323, 328
  luplot, 38–40
  max, 65, 242, 327, 386, 398–400, 406
  mesh, 51–53, 59, 62, 69, 70, 169, 411, 431
  meshc, 58–59
  meshgrid, 51–52, 59, 61, 64, 67, 70, 170,
    172, 174, 175, 431
  min, 65, 386, 398–400
  movie, 69–70
  namelengthmax, 17
  num2str, 17, 81
  ode113, 227
  ode15s, 227
  ode23, 226–228, 230, 235–239, 241, 243
  ode23S, 227
  ode23t, 227
  ode23tb, 227
  ode45, 227, 229, 231–232, 234
  odeset, 241, 243
  optimset, 321
  patch, 64
  path, 78
  pause, 397, 398, 403, 412
  plot, 25–28, 30, 35–38, 42–49, 73, 75, 80,
    223, 228, 230, 234–239, 241, 290, 293,
    324–326, 329, 371, 385, 396–397, 400,
    402–403, 405–406
  plot3, 74
  plotradio, 38
  plotrc, 49
  pol2cart, 59–60
INDEX 517

polar, 45
print, 50, 234, 326, 328
pwd, 30, 78
quad, 286–294
quadl, 286–294
quit, 16
quiver, 67
quiver3, 68, 90
rand, 20
return, 395, 401
rmpath, 78
roots, 321, 322, 324
rot90, 431, 433
rotate, 64
round, 17
save, 30
semilogx, 47–49
semilogy, 47
shading, 56, 62–64
single, 17
size, 239, 241
slice, 62–64
sort, 83
sprintf, 223
str2num, 17
subplot, 46, 75, 230, 371, 385, 400, 405–406
sum, 83, 283, 286
surf, 51, 55–56, 59, 62, 64, 170
surfc, 58–59
surf1, 58–59
text, 35, 37, 49, 76, 79, 239, 241, 329
trace, 23
transpose, 23, 35, 80, 284, 323, 324
trapz, 286
uint16, 17
uint32, 17
uint64, 17
uint8, 17
view, 53, 412, 431
warning, 325
who, 20, 36
whos, 20, 36, 229, 237
xlabel, 26–28, 43, 47, 52, 56, 67, 73, 79, 223, 228, 230, 290, 291, 294, 324, 326, 431
ylabel, 26–28, 43, 47, 52, 56, 67, 73, 79, 223, 228, 230, 290, 291, 294, 324, 326, 431
zeros, 20, 32, 62, 64, 169, 382, 383, 402, 406, 407, 426–428, 430, 432
zlabel, 52, 56, 79
MATLAB contact information, 506
MATLAB control structures
break, 82
function/end, 38–40, 226, 228, 230, 232, 235, 238, 287, 288, 291, 293, 323, 325, 328
if/else/elseif/end, 34, 395, 398, 401, 426, 428–429
return, 82
while/end, 33
MATLAB default directory, see current directory in MATLAB
MATLAB functions, see also MATLAB commands, MATLAB programs, 37–40
abs, 242
apollo, 243
built-in, 22
cos, 43, 45, 46, 59, 67, 70, 74, 75, 174, 291, 293, 326, 328
sin, 170
cosh, 25, 46
dampharm, 243
damposc, 230, 243
decay, 228
drvnosc, 243
drvnpend, 243
eig, 41–42
exp, 44, 48, 59, 61, 172, 283–286, 325, 326
fix, 395, 396, 400, 402–403
fsolve, 341
gausint, 287
henon, 243
largamp, 243
drwnpend, 23
lineshape, 39
lineshape1, 40
lorenz, 243
lubisect, 322, 350
lugen_list, 170–172, 175, 176
luread_list, 176
max, 372
min, 372
moment, 288
onedshm, 243
planet, 232, 244
projectile, 325
qmlower, 327
qmshn, 238, 244
qmupper, 327
rings, 293
rossler, 244
sin, 22, 43, 44, 46, 50, 52, 67, 70, 74, 75, 170, 174, 325, 328
sinh, 46
spiralc, 291
sqrt, 22, 48, 61, 172, 283–286, 289, 324, 328
stdwaves, 235, 244
INDEX

MATLAB handles, see handles in MATLAB
MATLAB linestyle codes, 45
MATLAB M-files, see MATLAB scripts, MATLAB functions
MATLAB objects, 28
MATLAB programs, see also MATLAB commands, MATLAB functions

- fdm1d.m, 370, 441
- fdmdiffus1d.m, 403–404, 445–446
- fdmlap2d.m, 409, 446–447
- fdmlap2d_plot.m, 411, 447–448
- fdmwave1d.m, 397–398, 444–445
- fem1d.m, 383, 442–443
- fem2d.m, 430, 453–455
- lugen_list.m, 183–184
- lureen_list.m, 185

- MATLAB properties, 18–19, 26–28, 42, 43
- AbsTol, 242, 243
- Box, 35
- ColorMap, 77
- ColorOrder, 36, 80
- DefaultAxesColorOrder, 80
- DefaultAxesLineStyleOrder, 80
- EdgeColor, 52, 62, 64, 69, 411, 431
- FaceColor, 64
- FixedColors, 77
- FontAngle, 71
- FontName, 71–72
- FontUnits, 71
- FontWeight, 71
- InitialStep, 242
-LineStyle, 44, 47–49, 223, 228, 230, 236
- Marker, 47, 49
- MaxStep, 242
- MenuBar, 74
- Name, 74, 75
- NextPlot, 35
- Position, 74, 75
- RelTol, 242, 243
- ToIX, 321
- Visible, 62, 63, 79

- XAxisLocation, 80
- XColor, 79
- XData, 64
- XGrid, 29, 30
- XLim, 28–30, 234, 325
- YAxisLocation, 80
- YColor, 79
- YData, 64
- YGrid, 29, 30
- ZColor, 79
- ZData, 64

- MATLAB scripts, 37–38
- MATLAB special symbols, 72
- ! (spawning), 16
- \n (new line), 30, 31, 81, 223, 408, 410
- \ (transposition), 69, 431
- ’...’ (enclose strings), 18, 19
- * (multiplication), 22, 23, 32
- + (addition), 22
- - (subtraction), 22
- .* (multiplication), 22, 32, 44, 50, 52, 59, 61, 67, 70, 170, 172, 174, 175, 289
- ./ (division), 22, 23, 39, 40, 48, 50, 52, 170
- .^ (exponentiation), 22, 43, 48, 51, 52, 59, 61, 73, 90, 170, 172, 175, 283, 284, 286, 289
- / (division), 22, 24, 339
- : (range), 21, 25, 51–52, 62, 174, 175, 323, 325, 326, 328, 369, 372
- ; (separate rows in array), 19
- ; (suppress output), 16, 19, 20, 23–25, 38, 44, 47–48, 52, 61, 64, 223, 228, 229, 232, 283, 285, 286, 290, 323–327, 329
- = (assignment), 17
- == (equality), 398, 403, 410
- % (comments), 16
- %d (format), 30–31, 408, 410
- %e (format), 31
- %f (format), 31, 223
- \ (division), 24, 339, 370, 383, 430, 432
- ^ (exponentiation), 22, 39, 40
- ~ (logical not), 33
- ~= (logical not equal), 429

for expressing conditions, 33

MATLAB system variables
- ans, 18
- i (√−1), 19
- j (√−1), 19
- NaN, 326, 328
- path, 78
- pi (π), 22, 48, 50, 52, 67, 70, 75, 170, 172, 174, 283–286, 289, 325, 326
INDEX

in MATLAB, 19–23
in MAXIMA, 113–115
matrix multiplication
in MATLAB, 23
in MAXIMA, 114
max-min problems, 306
MAXIMA, 1
MAXIMA command files, see also MAXIMA
commands, MAXIMA functions, 119–123
crossdot, 121, 123, 152, 267
MAXvectcart, 115–117, 150
MAXvectcyl, 151
MAXvectsph, 151
testmultiply, 120–121
MAXIMA commands, see also MAXIMA
command files, MAXIMA functions
abs, 104
algsys, 341
allroots, 330
append, 138
appendfile, 131
apply, 140
apply1, 135, 136
apply1b, 135
apply2, 135
apropos, 141
assume, 141, 264–266, 268, 272
atvalue, 201, 207, 208
batch, 119–123
batchload, 119–123
bc2, 201, 209
bf_find_root, 330, 331
bfallroots, 330
bfloat, 101
cabs, 104
carg, 104
closefile, 131
coeffmatrix, 209
collectterms, 202–203, 211, 213
combine, 142
compile, 122, 296, 297
compile_file, 122
conjugate, 104
countour_plot, 127–130
contrib_ode, 201
cross, 121–122
curlcart, 116, 150
curlycyl, 151
curlsp, 151
declare, 141, 210, 272, 274
defmatch, 137
defrule, 135, 136
demo, 96, 119, 121
demoivre, 105, 209
denom, 105
depends, 110, 200–202, 204, 209
describe, 96
desolve, 201, 207, 208
determinant, 114, 209
diff, 108, 110–112, 125, 144, 147, 202–205,
207–209, 211, 314
dispfun, 141
distrib, 142, 211
divcart, 116, 150
divcyl, 151
divsph, 151
dot, 121–122
eigenvalues, 115
eigenvectors, 115
eivals, 115
eivects, 115
ev, 103, 106, 112, 203, 205, 206, 209, 210,
212, 213, 271, 332
example, 96, 119
expand, 102, 103, 106, 107, 112, 142,
202–204, 270, 271, 316–317, 330
exponentialize, 105
factor, 102, 108, 205, 212, 268, 272, 274,
316
find_root, 270, 330, 331, 334
float, 101, 102, 107, 265, 270–272, 295,
314, 331
forget, 141
get_plot_option, 123
gradcart, 116, 150
gradcyl, 151
gradsp, 151
ic1, 201, 204
ic2, 111, 113, 201, 202
ilt, 110, 149, 201
imagpart, 104, 314
invert, 114
is, 203
kill, 95, 102, 103, 105, 106, 108–110, 113,
115, 130, 136
lapcart, 116, 150
lapcyl, 151
laplace, 109, 110, 149, 201
lapsp, 151
lhs, 203, 205
limit, 205, 207, 264, 317
linsolve, 339
load, 114, 119–123, 202, 267, 331, 332
loadfile, 141
logcontract, 205, 267
makelist, 106, 245, 296, 297, 333, 334
map, 141, 314, 315
matchdeclare, 134–137
matrix, 113, 114
mode_declare, 123, 297
multthru, 142
newton, 330–332, 341
nroots, 330, 331
num, 105
ode2, 111, 113, 200, 202, 204, 209
INDEX

part, 102–103, 105, 142, 204, 205, 207, 212, 245, 264, 268, 270, 314, 316, 317
parfrac, 103, 149
pickapart, 142
playback, 131
plot2d, 124–126, 139, 205, 206, 208, 245, 269, 274, 296, 297, 315, 333, 335
plot3d, 126–127, 140
polarform, 105
print, 102, 107, 212
product, 141
quad_qag, 294–297
quad_qagi, 294
quad_qags, 294
radcan, 142, 148, 205
ratsimp, 108, 142, 205
realpart, 104, 314
realroots, 330, 331
rectform, 105
remfunction, 100, 141
remove, 141
remvalue, 100, 107, 141
rest, 334
rhs, 202, 203, 205, 207, 209, 210, 314, 315
rk, 244–245
save, 131, 141
set_plot_option, 123
solve, 106, 107, 204, 205, 209, 212, 264, 270, 313–317, 339, 341
stringout, 131, 141
subst, 103, 111–112, 142, 202–203, 205, 264, 267, 268, 271
substpart, 103, 142
sum, 141, 211, 274
taylor, 109, 204, 213, 264, 267, 269, 271, 272, 316, 317
tellsimp, 136
tellsimpafter, 136
tex, 132–133
tracematrix, 114
translate, 122
translate_file, 122
transpose, 114
trigexpansion, 105, 106
trigreduce, 105
trigsimp, 106, 204, 266, 268
ueivects, 115
uniteigenvectors, 115
writefile, 131
wxplot2d, 124
wxplot3d, 126–127
xthru, 142
zeromatric, 148
MAXIMA contact information, 506
MAXIMA control structures
block(...), 118, 332
for/step/unless/do, 117–118
for/step/while/do, 117–118
for/thru/step/do, 102, 106, 107, 117, 212, 213, 296, 297, 332, 334, 335
if/then/elseif/.../else, 118–119
unless/do, 119
while/do, 119
MAXIMA examples, 96
MAXIMA functions, see also MAXIMA commands, MAXIMA command files, 97–98, 100
asinh, 266
atan2, 104, 105
cosh, 104
cot, 334
exp, 105, 108, 109, 123, 205, 295, 296, 315, 331, 332
factorial, 101
map, 205
sin, 100, 104–106, 109, 112, 113, 123, 126, 135, 136, 139, 140, 263, 266, 268, 315, 332, 334
sqrt, 101, 107, 111, 264–266, 295–297, 315, 334
tan, 113
MAXIMA keywords, 97
epsabs, 295
epsrel, 295
limit, 295
MAXIMA logical expressions, see logical expressions in MAXIMA
MAXIMA manuals, 93
MAXIMA notebooks, see MAXIMA workbooks
MAXIMA option variables, 97
algebraic, 142
file_search.demo, 96
file_search.lisp, 119–120, 138–139
file_search.maxima, 119–120, 138–139
file_type.lisp, 119–120
file.type_maxima, 119–120
fpprec, 101
fpprintprec, 101
maxima_tempdir, 123, 131, 132, 137–138
maxima_userdir, 131, 132, 137–138
method, 204
MAXIMA packages
functs, 114
newton1, 330–332
orthopoly, 144, 147
MAXIMA plot options, 97
black, 269
color, 124–127, 206, 208, 245, 269, 274, 296, 297, 315, 333, 335
color_bar, 127
discrete, 245, 296, 297, 333, 335
gnuplot_preamble, 127–130
grid, 126, 140
grid2d, 124–125, 296, 297
INDEX

label, 124, 129, 139
label_alignment, 139
label_orientation, 139
legend, 125–130, 140, 206, 208, 245, 274, 315, 335
lines, 125, 206, 208, 245, 269, 274, 296, 297, 315, 333, 335
mesh_lines_color, 127
nticks, 125
palette, 126
parametric, 126
plot_format, 123
same_xy, 126, 140
style, 125, 206, 208, 245, 269, 274, 296, 297, 315, 333, 335
t, 140
title, 124–130, 296, 297
value, 127
xlabel, 124–126, 206, 208, 245, 269, 274, 296, 297, 315, 333, 335
xtics, 140
y, 124–126, 140
ylabel, 124–126, 206, 208, 245, 269, 274, 296, 297, 315, 333, 335
ytics, 140
z, 140

MAXIMA primer, 93

MAXIMA special symbols, 97

! (factorial), 101
!! (double factorial), 101
' (suppress evaluation), 110–112, 202, 204, 207–209
* (element-by-element product), 114
* (multiplication), 98
** (exponentiation), 98
+ (addition), 98, 101
- (subtraction), 98
. (matrix product), 114
/ (division), 98, 101
/*...*/ (comments), 116, 121
: (assignment), 99, 106
:= (function definition), 100, 123
; (display output), 94
? (in-line help), 95
$ (suppress output), 94, 101–103, 105
% (previous statement), 107, 108
%c (ODE constant), 111, 200–201
%e (e), 97, 110
%gamma (γ), 97
%i (√−1), 97, 104
%k1 (ODE constant), 111, 200–201, 209
%k2 (ODE constant), 111, 200–201, 209
%pi (π), 97, 106, 126, 210, 295–297, 332
~ (exponentiation), 98, 99, 102, 106, 108
inf (∞), 97, 264
minf (−∞), 97, 264

MAXIMA statements
termination of, 94

MAXIMA system variables, 97

features, 141
functions, 141
inchar, 94
infolists, 141
intfactor, 200
labels, 141
maxima_userdir, 139
method, 111, 200, 202
multiplicities, 106, 330, 331
rootsepsilon, 330
values, 141
wxplot_size, 124
yp, 200

MAXIMA workbooks, 133–134
Maxwell-Boltzmann distribution, 301
method of bisection, see root finding
method of Frobenius, 211–213
method of least squares, see least squares fitting
midpoint method, 218
midpoint rule, see quadrature
MiKTeX, 1
MiKTeX contact information, 506
ML (left mouse button), 9
MM (middle mouse button), 9
moment of inertia
definition of, 257
do cardboard, 299
do of semicircular plate, 257, 265, 288–289
motion
do of membrane, 69–71, 91
do of string, 91
MR (right mouse button), 9
MUDPACK, 1
MUDPACK contact information, 506
Müller’s method, see root finding
multi-way selection
in pseudocode, 159
multistep methods
for ODEs, 222

N

NAG library contact information, 506

named common
in MATLAB, see global variables

natural frequencies
of spherical cavity, 344
of stiff bar, 343, 344
of three-mass system, 342
of two-mass system, 307–308, 315, 323–324
Navier-Stokes’ equations, 359
Newton’s method, see root finding
Newton’s second law, 255, 256, 307, 309
in fluids, 361–363
Newtonian fluid, 363
node numbers
global, 374
local, 374
nodes, 351
INDEX

non-viscous fluid flow, 363
normal distribution, see Gaussian distribution
normal modes
  of membrane, 50–58
  of string, 91
normal stress
  in fluids, 361–363
numerical approximation
  for first derivative, 366
  for second derivative, 367
numerical integration, see quadrature
Numerical Recipes, 1
NUMERICAL RECIPES contact information, 507
nybble, 5

O

object code, 154
object-image relations, 85
object-oriented programming, 28
octal number system, 5
OCTAVE, 1
OCTAVE contact information, 507
on-line help
  in MATLAB, 36–37
  in MAXIMA, 95–96
ordinary differential equations
  with MATLAB, 222–244
  with MAXIMA, 110–113, 200–213, 244–245
orthogonal functions, 262–263
OzTEX contact information, 507

P

page previewer
  in \LaTeX, 488–489
partial fractions, 103, 149, 247
passwords, see Local Guide
path (in directory tree), 8
path specification
  in MATLAB, 77
  in MAXIMA, 137
PDEs
  classification of, 358–359
PDF files, 461–462
pdffcrop, 484–485
pdflatex, 461, 478, 483–484
pdftops, 485
pendulum, simple, see simple pendulum
pivotal condensation, 338
pivoting, 337
placing text in graphics
  in MATLAB, see MATLAB commands, text
  in MAXIMA, see MAXIMA plot options, label
Planck radiation law, 87
Poisson’s equation, 358
polar coordinates
  in MATLAB, 45, 59–60
  in MAXIMA, 105, 117
population biology, 192, 204–206, 252
PostScript, 50
preallocation of arrays
  in MATLAB, 32, 62
predictor-corrector methods, 216
program, 153
projectile motion, see trajectories
properties
  in MATLAB, see MATLAB properties
proprietary software, 14
protected variables
  in MATLAB, see MATLAB special symbols, MATLAB system variables
  in MAXIMA, see MAXIMA special symbols, MAXIMA system variables
ps2pdf, 462, 479, 484–485
pseudocode comments, 159, 162
pseudocode control structures, 153, 158–162
CASE-END_CASE, 159
EXECUTE, 161
FOR-THRU-DO, 160–161, 166–168
FOR-THRU-DO-STEP, 160
IF-THEN-ELSE-END_IF, 3, 4, 159, 163, 164
LOOP-END_LOOP, 3, 4, 159, 162–164
REPEAT-UNTIL, 160
WHILE-DO, 160
pseudocode directives
  DIMENSION, 155, 166, 168
pseudocommands
  ATTACH, 168
  CLOSE, 168
  READ, 156
  WRITE, 156, 166, 168
PYTHON, 1
PYTHON scripts
  ExtractFileName.py, 504
Q

Q of resonant circuit, 88
quadratic formula, 107, 312, 313, 341, 343
quadrature, 273–283
  adaptive algorithms, 280
  assessing error in, 278–279, 283–286
  Gauss-Legendre, 283
  Gaussian, 280–283, 285–286, 301
  iterative algorithms, 280
  midpoint rule, 150, 275
  Newton-Cotes, 273–279
  recursive trapezoidal rule, 277
  Romberg integration, 278
  roundoff error, see roundoff error
  Simpson’s rule, 149, 150, 276, 278, 282, 284, 301
  trapezoidal rule, 275, 283
INDEX

truncation error, see truncation error
with MATLAB, 283–294
with MAXIMA, 294–297
quality factor, see Q of resonant circuit
quantum energy levels, see energy levels
quantum probabilities, 262
  in hydrogen atom, 61, 88, 171–172, 181, 299, 303
  of harmonic oscillator, 262, 289, 296, 301
quantum scattering, 145
quantum wave function, see wave function

R

radius of gyration
  definition of, 257
  of cardioid, 299
  of semicircular plate, 265
railroad problem, 345
range of projectiles, see trajectories
Rayleigh criterion, 345
reaction kinetics, 252
reading data
  from files
    in MATLAB, 34–36, 176
    in pseudocode, 169
  from keyboard
    in MATLAB, 81
    in pseudocode, 156
recasting
  algorithms, 177
  control structures, 177
recursion relationship, 211
reflection at dielectric film, 86
relativity
  mass/momentum/energy, 86
  motion under constant force, 264
relaxation, 165–166
residual, 320
resonance
  in LRC circuit, 88, 194, 248
restoring initial state
  in MATLAB, 79
  in MAXIMA, 95, 139
returning to command level
  in MATLAB, 16
  in MAXIMA, 94
returning to operating system
  from MATLAB, 16
  from MAXIMA, 94
ReVTeX, 493
Richardson extrapolation, 277, 284
Ritz variational method, 373
Rodriques’ formula, 143, 146
Romberg integration, see quadrature
root finding
  accuracy in, 320–321
by bisection, 318, 322
by Laguerre’s method, 320
by Müller’s method, 319
by Newton’s method, 319, 340–341, 343
by secant method, 319
by trial and error, 339–340
in MATLAB, 321–329, 339, 341
in MAXIMA, 313–317, 329–335, 341
roundoff error, 368–369, 373, 395
  in numerical quadrature, 278
  in ODEs, 219
rubber lengths in Î²Î±Î³ÎµÎµÎµ-Î»X, 463
Ruchardt’s experiment, 348
Runge-Kutta methods, 217–219
running program, see Local Guide

S

saving files, see writing data to files
scalar fields
  three-dimensional, 171–172
    in MATLAB, 172
  two-dimensional, 170–171
    in MATLAB, 170–171
search path
  in MATLAB, 15, 78
  in MAXIMA, 120, 138–139
secant method, see root finding
self-adjoint ODE, 365
separating statements on one line
  in MATLAB, 16, 33, 46
  in MAXIMA, 94
  in pseudocode, 158
series solutions to ODEs, 211–213
shape functions, 373–375
shear stress
  in fluids, 361–363
shooting method, 234
sigmoid curve, 192, 207
simple pendulum, 250, 258, 271–272, 303
Simpson’s rule, see quadrature
simultaneous equations
  by direct methods, 338
  by iterative methods, 338
    linear, 336–339
    nonlinear, 339–341
singular equations, 336
singular value decomposition, 338
small angle approximation
  in MAXIMA, 135–136
solving Laplace’s equation, see Laplace’s equation
solving linear equations, 336–339
  in MATLAB, 24–25
  in MAXIMA, 107
solving ODEs
  with MATLAB, 222–244
with MAXIMA, 200–213, 244–245
sound waves in fluids, 364–365
source code, 154
spawning to operating system
from MATLAB, 16
special symbols
in MATLAB, see MATLAB special symbols
in MAXIMA, see MAXIMA special symbols
speed of sound, 364
spherical coordinates
in MAXIMA, 151
spurious roots, 305, 311, 328, 334
stable equilibrium, 306
standing waves
in cube, 89, 182
in string, 197–198, 209–210, 234–237, 249
Stark effect, 84, 147, 343
startup files, see initialization files
structure of data files, 166–167
subroutines
in MATLAB, see MATLAB functions,
MATLAB programs
in MAXIMA, see MAXIMA functions
suppressing evaluation
in MAXIMA, 111–112
suppressing output
in MATLAB, 16, 19, 20
in MAXIMA, 94

T

tables
in L\LaTeX, 472–474
Taylor series
for solving ODEs, 217–221
for solving simultaneous equations, 341
in MAXIMA, 109, 143, 149
termination of statements
in MATLAB, 16
in MAXIMA, 94
in pseudocode, 158
\LaTeX, 457, 465
\LaTeX contact information, 457, 507
\LaTeX output
from MAXIMA, see \LaTeX output
texdoc, 477, 498
\TeXlive, 1
\TeXlive contact information, 507
text editors, see Local Guide
text in graphics, see placing text in graphics
TGIF, 1
TGIF contact information, 508
thermal equilibrium, 88
thin lens equation, 85
tikz, 477–480
tikz commands
\definecolor, 478
\draw, 478–479
\foreach, 478, 479
\node, 478, 479
\x, 478, 479
trajectories, 255
double pendulum, 247
double pendulum, 247
earth falling into sun, 298, 302
exponentially decaying force, 298
in central forces, 195–197, 231–234, 250, 251
in constant magnetic field, 73, 140, 181
in Coulomb field, 248
in crossed E and B, 251
in viscous medium, 188–191, 202–204, 248,
308–309, 315–317, 324–327, 332–333, 342
simple harmonic oscillator, 298
under constant force, 264, 297, 298
transmission at dielectric film, 86
transpose
in MATLAB, 23
in MAXIMA, 114, 121
trapezoidal rule, see quadrature
tree-structured directories, 8
tridiagonal systems, 338, 348
truncation error
global, 221
in midpoint formula, 279
in numerical quadrature, 278
in ODEs, 219, 223–225
in Simpson’s rule, 279
in trapezoidal rule, 279, 283
per step, 220
truth table, 157
turning points, 262, 289, 301, 306, 313–314,
322–323, 331–332, 344, 345
two’s complement notation, 6
two-way selection
in MATLAB, 34
in MAXIMA, 118–119
in pseudocode, 159

U

UNIX, 1
UNIX commands
aspell, 489–490
cleanbak, 498
clean\TeX, 498
dvipdfm, 462
dvips, 461
dvipdfmx, 462
dvipdfmtex, 462
dvi, 462
dvi, 462
dvips, 461
pdftops, 485
ps2pdf, 462
texdoc, 477
xdvi, 462
UNIX shell scripts
INDEX

ceps2pdf, 485, 503

cleanbak, 498

cleanTEX, 498

cps2pdf, 485, 503

dfileeps, 485, 503

dfileps, 485, 503

UNIX tutorial, see Local Guide

unstable equilibrium, 306

user accounts, see Local Guide

user-defined commands, see command files

user-defined functions

  in MATLAB, see MATLAB functions

  in MAXIMA, see MAXIMA functions

V

Van der Pol oscillator, 250

variable names

  in MATLAB, 17

  in MAXIMA, 97

  in pseudocode, 154

vector arrays

  three-dimensional, 175–176

  two-dimensional, 172–175

vector calculus

  with MAXIMA, 115–117

vector fields

  three-dimensional

    in MATLAB, 175–176

  two-dimensional

    in MATLAB, 174–175

vector identities, 152

velocity field

  of rotating object, 182

vertices, 351

vim, 1

viscosity, 363

viscous damping, 189, 192

W

wave equation

  in 1D, 353–355

  in MATLAB, 395–400

  in 2D, 387–389

  in 3D, 355–356

  wave function, 198–199, 237–241, 262, 300, 301

  weighting functions, 376

Wheatstone bridge, 144

Windows, 1

Windows batch files

  ceps2pdf.bat, 485, 502

  cleanbak.bat, 498

  cleanTEX.bat, 498

  cps2pdf.bat, 485, 502

  rdfileeps.bat, 485, 502

  rdfileps.bat, 485, 502

Windows commands

  cleanbak, 498

  cleanTEX, 498

  dir, 485

  dvipdfm, 462

  dvips, 461

  latex, 460, 461

  pdfcrop, 484–485

  pdflatex, 461

  pdftops, 485

  print, 462

  ps2pdf, 462

  texdoc, 477

  yap, 462

Winedt, 1

Winedt contact information, 508

writing data to files, 181, 182

  in MATLAB, 30–31, 170–172, 174–176

  in pseudocode, 166–169

writing data to screen

  in MATLAB, 19–21, 81

  in pseudocode, 156

X

xdvi, 462, 488–489

Xemacs, 1

Xemacs contact information, 508

Xv contact information, 508

Y

yap, 462, 488–489